Feature-centric ranking algorithms for georeferenced video search

Holger Fritze
Institute for Geoinformatics
Münster, Germany
h.fritze@uni-muenster.de

Auriol Degbelo
Institute for Geoinformatics
Münster, Germany
degbelo@uni-muenster.de

Tobias Brüggentisch
Institute for Geoinformatics
Münster, Germany
t.brueggentisch@uni-muenster.de

Christian Kray
Institute for Geoinformatics
Münster, Germany
c.kray@uni-muenster.de

ABSTRACT

While it is commonplace to retrieve photos showing a particular feature (e.g. through tools such as Google Pictures or Bing Images), spatial approaches for retrieving videos showing a particular feature (e.g. a building) have yet to be established. This article proposes five ranking algorithms to query georeferenced videos for a specific feature based on the videos' spatio-temporal metadata. 12 relevance criteria for feature-centric video ranking were compiled from a focus group discussion. From these, four criteria have been selected for implementation: "Feature Depiction", "Feature Illumination", "Feature Visibility Duration", and "Distance to Feature". These criteria were implemented in five algorithms and evaluated regarding efficiency and user perceived plausibility. The evaluation suggests that the "Feature Visibility Duration" of the video's viewshed with the queried feature geometry offers a good trade-off between computationally performant and cognitive plausible ranking. The obtained results are relevant to user-centered approaches for interacting with georeferenced videos.

CCS CONCEPTS

•Information systems →Information retrieval; •Human-centered computing →*Human computer interaction (HCI);*

KEYWORDS

georeferenced video, video search, video ranking

ACM Reference format:
Holger Fritze, Auriol Degbelo, Tobias Brüggentisch, and Christian Kray. 2017. Feature-centric ranking algorithms for georeferenced video search. In *Proceedings of SIGSPATIAL'17: 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Los Angeles Area, CA, USA, November 7–10, 2017,* 10 pages.
DOI: 10.1145/3139958.3139976

SIGSPATIAL'17: 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Los Angeles Area, CA, USA

DOI: 10.1145/3139958.3139976

1 INTRODUCTION

Thanks to technological advances, video creation and sharing have become commonplace. Online hosting platforms such as Youtube or Vimeo enable hundreds of millions of users to share and view video content from various sources, which is captured by increasingly more mobile devices. Scientific work is currently underway to produce efficient techniques to search for videos by spatial features and their characteristics. For example, Lu et al. [32] introduced a dataset which can be used to advance research on spatio-temporal video search. Emrich et al. [7] presented a system that enables retrieving georeferenced videos by a user-defined trajectory. Ay et al. [1] presented a prototypical implementation of a web-based, georeferenced video search engine. On a conceptual level, Yin et al. [45] distinguished between content-based and context-based techniques for georeferenced video retrieval. Content-based retrieval methods focus on visual features in the video and need pre-processing so that visual information is extracted, coded, and stored. Context-based techniques rely on metadata (e.g. camera location, camera orientation) for georeferenced video retrieval, and have no need for pre-processing.

Current search engines such as Google or Bing support feature-centric retrieval, i.e. they help to retrieve pictures or videos which show a particular feature (e.g. Eiffel Tower). Still, supporting users in formulating spatial queries about videos (e.g. retrieve videos showing the Eiffel tower whose spatial location are within a bounding box) is currently limited. We do not know which (spatial) criteria matter the most to users when searching for videos which show a particular feature, nor do we know how to translate these criteria into useful ranking metrics for algorithms. This article explores these two questions. The work aims to take advantage of sensor metadata to provide a feature-centric retrieval of georeferenced videos. It is therefore context-based but strives to provide a ranking which reflects the characteristics of prominent landmarks in the video. The work is also user-centered because it takes into account criteria which people perceive as important regarding video relevance. The main contributions of the article are as follows:

- a set of 12 criteria to consider while developing feature-centric algorithms for georeferenced video search. These criteria were extracted from a focus group interview with six participants;
- five algorithms (and a web-based application) for feature-centric georeferenced video ranking. These algorithms take four selected criteria from the focus group interview into account,

namely degree of depiction of a particular feature in the video, the feature illumination, distance of video to the feature, and the video duration with respect to the feature;
- an assessment of the computational efficiency and the cognitive plausibility of the four criteria.

In the remainder of the article, the terms "feature", "object of interest" and "query object" are used interchangeably to denote the object shown in the video (e.g. building, tourist attraction) which is of interest to the user. The development of the feature-centric algorithm in this work involves three steps: First, an identification of the relevant criteria for feature-centric search from a user perspective; second, a proposal of algorithms which implement four of the criteria; and third an evaluation of the cognitive plausibility of the algorithms via a user study. Related work is briefly reviewed in Section 2 before presenting the focus group interview, and the criteria for feature-centric video ranking in Section 3. The algorithms suggested in this work are introduced in Section 4. Section 5 discusses the cognitive plausibility and computational efficiency of the criteria, Section 6 touches on the limitations of the work, and Section 7 concludes the article.

2 RELATED WORK

The ubiquitous availability of camera-equipped smartphones has not only established but further increased the popularity of sharing videos [37, 48]. Videos are recorded for a wide variety of occasions and purposes [2, 4, 40], turning online platforms, such as Youtube and Vimeo, into a central part of the mainstream media landscape [14]. However, searching and indexing video collections are challenging tasks [2]. This raises the need for comprehensive search algorithms to efficiently query large video datasets.

To query videos, search engines need to know about the video content. For videos showing real-world aspects, the question of *what* is seen in the video strongly relates to *where* the video has been recorded. To also provide a spatial context, search engines need to know the video recording location. In principle, this information can be obtained by explicit or implicit geographic tagging [15]. While the explicit approach relies on embedded sensor data and records the location information during the content creation, the implicit approach produces the spatial metadata by post-hoc parsing of the textual descriptions. If geospatial metadata is explicitly available for a video, it is often referred to as spatial video [28, 40], geospatial video [42, 46], or Full Motion Video [34, 36]. In the context of social networks, the term *geo-social (multi-)media* is also used for photos and videos to emphasize a focus on explicit spatial metadata [9, 15, 35].

Search engines mainly rely on the implicit approach and use annotations or content analyses for indexing and retrieval tasks [23, 33]. The textual information (e.g. tags, titles, and descriptions) is contingent on the manual input provided by the users. Hence, the quality of search queries is strongly dependent on the quality of contextual information provided by the creators. Those annotations are subjective and lack precision [39], and they are often insufficient for meaningful and accurate search results [45].

Content-based methods also follow the implicit approach and utilize techniques such as visual feature extraction and segmentation to evaluate the user queries [6]. Several major efforts have been made in the field of content-based video retrieval techniques [16, 43]. These approaches directly process the video content and are therefore less dependent on the provision of metadata. However, they are also computationally expensive [39], which makes the utilization of associated metadata in many cases a more feasible approach [2, 45].

The contextual metadata of the video can be used to infer its location [18, 29] and to turn an implicit location information such as a city name or a landmark into explicit coordinates. However, with an achieved precision at a city level (10 km), a derived geotag can only inform about the video in its entirety and provide a rough location estimation. A single geotag is in many cases not sufficient to describe a video properly. It may remain unclear if the geotag refers to the start or end location of the video, to the recorded landmark, or if the geotag indicates to the recorded neighborhood or city of the video. Accordingly, the relation between the video contents and their cartographic representations remains very loose [37].

To fully describe the spatial properties of a video recording the metadata need to explicitly model the viewable scenes of the video. This includes among others an explicit model of the viewshed of the camera with the change of properties over time. Ay et al. [4] presented a viewable scene model consisting of a set of Field of View Scenes (FOVScenes). Each scene is defined as

$$FOVScene(P, \vec{d}, \theta, R) \tag{1}$$

and described by P: camera location as geotag ($< Lat, Lon >$), θ: the viewable angle, $\vec{d}$: the camera direction vector, and R: the visible distance. The resulting geometry forms a pie-slice-shaped area. An alternative model was presented by Lewis et al. [28]. Their approach, called Viewpoint, extends the OGC ViewCone model[1] and represents the camera FOV as 3D shape with a near depth-of-field as minimum distance. The view cone has the shape of a rotated and truncated pyramid. The FOV geometries are derived from logged GPS information.

Videos enriched with a geospatial description of the viewable scenes can be queried and analyzed spatially by retrieval systems. Scholars have developed different applications which are now briefly discussed. In 2003, Kim et al. [21] presented a concept of interactive geographic videos called GeoVideo. They introduced the term *MediaGIS* for GIS software that provides tools for integrating multimedia and spatial information. The system relates video contents to objects within a 3D virtual scene and allows three ways of interaction between the video content and the corresponding geography: (1) geography-to-video interaction, (2) video-to-geography interaction, and (3) a mutual interaction of video and geography. In this application, the geography is depicted by a 3D virtual environment. In 2009, Ay et al. [3] demonstrated a prototype of a georeferenced video search engine (GRVS), which utilizes the aforementioned FOVScene as estimation model. By sliding through the video, a map interface depicts the corresponding location, orientation, and viewshed information. Similar to GeoVideo, this application also does not incorporate information about features that lie within a FOV. Zhang et al. [49] used the jointly recorded geosensor stream to register the video within georeferenced 3D models. Similarly,

[1] OGC Geo-Video Web Service: http://portal.opengeospatial.org/files/?artifact_id=12899

Seo et al. [37] extracted keyframes and placed them in their correct geographic locations on a map instead of the location where the videos were captured. For objects close to one another, this approach was lacking the ability to distinguish them and to do a correct placement. With the advent of User Generated Videos (UGVs) Zhang and Zimmermann [50] developed an application to generate a video summary for path queries. The application combines multiple videos, which have been recorded with specially designed apps for smartphones. Kim et al. [20] used metadata from smartphone sensors to automatically geotag (and later query) indoor videos.

The increasing amount of georeferenced videos stresses the need to index them for efficient retrieval. Research on spatial indexing was primarily conducted by the computer vision community, with approaches on content-based video indexing [11] and spatial indexing for videos [38]. However, these studies focused on indexing the relative location of objects in the video rather than the geographic placement of the video as such. Lu et al. [31] introduced R-tree-based indexes for location, orientation, and distance information of Fields of Views (FOVs). Gilboa-Solomon et al. [8] looked into efficient storage of time-stamped geographical tags in a spatial database. Kim et al. [22] used the FOVScene model by Ay et al. [4] and improved the concept of a Minimum Bounding Rectangle (MBR). They introduced *GeoSearch* as data structure, which uses Minimum Bounding Tilted Rectangless (MBTRs) to merge multiple similar FOVScenes into a larger representation. This data structure provides the basis for the *GeoTree* index [24] and the extended *GeoVideoIndex* Lee et al. [26]. While GeoVideoIndex can efficiently exclude superfluous scenes and substantially reduces the index size, the queries are confined to point and range queries.

In many situations, and particularly in urban environments, the view is limited and occluded. The viewshed is not equally distributed but follows and aligns to visual axes. Although models such as the FOVScene provide well-suited approximations for indexing approaches, retrieval systems need to incorporate the spatial settings of the environment for more sophisticated queries. Objects of interest may by occluded by other in-between objects, and the direct view may be limited. Or the object of interest is only visible on the edge of the viewshed for a short moment. Such issues show the need for more feature-centric approaches. If multiple videos show the same object of interest, there is also the need to rank the videos according to the relevance towards this object. Shen et al. [39] have generated descriptive tags about visible objects in the video scenes and ranked them according to their relevance. In Li et al. [30] textual and visual descriptions associated with videos were used to rank videos.

Searching and querying videos upon spatial properties require an explicit or implicit geographic tagging [15]. While the implicit tagging uses (manually) provided metadata to derive rough location information, the explicit approach relies on recorded sensor readings, which quantify spatial properties over time. We reviewed different modeling approaches to describe the viewshed of a video and its changes over time. To allow searching videos for a particular feature and prioritize them, the videos need to be ranked according to their relevance to the feature. Ay et al. [2] introduced three ranking algorithms that consider the spatial, temporal and combined spatio-temporal properties of georeferenced video clips. In contrast to descriptions and annotation-based approaches, these three metrics were solely based on the recorded explicit spatio-temporal metadata. Each metric was also evaluated with respect to a specific querying feature. The metrics further defined relevance scores for the *TotalOverlapArea*, *OverlapDuration*, and *SummedAreaOfOverlapRegions*. However, their work did not consider an occluded view. In this work, we have incorporated these metrics and used them as starting point for the exploration and development of further metrics.

3 CRITERIA FOR FEATURE-CENTRIC VIDEO RANKING

Ranking videos according to their relevance for a particular feature requires a set of characteristics against which the videos can be assessed. To understand the user's view about which characteristics matter the most, we conducted a focus group study with six participants. We preferred this qualitative research method over other methods as it allows to obtain multiple viewpoints and in-depth conversations and discussions [5, 25]. The group consisted of four men and two women with age between 18 and 27. The highest educational levels range from secondary school level 1 over A-level to a Bachelor's degree. All participants indicated low to moderate experience in the creation of geotagged media, and in the uploading of videos to video hosting platforms. The familiarity of the participants regarding the use of geotagged media and videos was higher. As a result, the participants are more representative of users who "consume" videos than of users who produce them.

For this focus group as well as the implementation and evaluation, we used an available dataset of georeferenced videos from the city of Singapore. As this dataset has also been used by Ay et al. [2], this allowed a comparison of the results. None of the participants has ever been to Singapore. The whole focus group discussion was tape-recorded and transcribed. Since the discussion took place in German, the quotes mentioned later in this section were translated from German into English. After assessing the participants' demographics, we provided an introduction into georeferenced videos and briefly demonstrated a retrieval system by Seo et al. [37] called GeoVid[2]. This application performs spatial queries, extracts single video keyframes by the FOV and allows placing them on the map. The demonstration served as illustration of how spatial videos can be queried. The discussion between the participants was moderated by one of the authors, and was guided by the following three questions:

(1) Which criteria/properties do you think should be considered to determine a video's relevance for a particular feature?
(2) Which (spatial) criteria/properties should be considered to determine a video's relevance for a particular feature in the context of georeferenced videos?
(3) How would you define the notion of relevance for a given video with respect to a particular feature in general?

After answering the three questions in turn, and discussing together the criteria which emerged, the participants were asked to perform a video comparison. We randomly selected nine videos showing the Marina Bay Sands Hotel in Singapore and created seven

[2]GeoVid WebViewer: http://api.geovid.org/v1.0/web/viewer

video pairs. The hotel was selected due to its size and prominent shape. For each video pair, the participants were asked to explain which video they would consider more relevant with respect to the Marina Bay Sands, and the criteria which they considered important. After each participant had shared her individual opinion, the group discussed the meanings of the emerging criteria and created the final set of criteria (i.e. added new criteria or removed criteria no longer considered important). This procedure was repeated for the seven video pairs.

At last, the FOVScene model from Ay et al. [4] was explained, and the participants learned about the properties to estimate the viewshed. The three guiding questions were asked again and answered a second time. The participants subsequently finalized the compiled relevance criteria collection and provided a short explanation of each criterion. To turn the criteria collection into a prioritized list sorted by importance, each participant received 40 marking points and was asked to distribute them across all identified criteria. Table 1 lists the 12 compiled criteria, their descriptions and respective scores. According to the participants, the "Complete Depiction" of the query feature is the most important characteristic, followed by the "Camera Work" and the "Focus / Sharpness" of the video. Some participants stated that the query feature should not be occluded by other features:

> *"I think the hotel has to be depicted completely. Otherwise, the video is not suitable for us at all." (P1)*[3]

> *"The view should not be blocked by other objects." (P5)*[3]

These quotes show how important the focus group considers an unoccluded view of the queried feature. The table also indicates that the video quality (i.e. the sharpness and camera work) is of major importance for the participants. The fourth characteristic directly relates to the illumination conditions of the query feature. The participants agreed that videos recorded at daytime are more relevant than those recorded at night. Regarding the temporal dimension, the participants argued that the video duration should be evaluated relative to the amount of time the query feature is visible in them. Hence, a short video showing the query feature over the whole duration would score better than a long video showing the feature for just a short moment while panning around. According to the focus group, a video's relevance increases with the relative time the query feature is shown.

The group also specified that the dimension of the feature also needs to be apparent when viewing a video. They claimed that videos showing only the queried feature itself make it difficult to determine the feature's size compared to other features. However, this characteristic was not considered as one of the most important ones. Similarly, the distance between the camera and the feature, as well as the camera height were mentioned, but considered less important among the specified criteria. This list can not be considered complete. The aspect that features located closer to the center of a scene receive more attention has been documented in the literature [17, 41], but was not included as a criterion by the focus group, although mentioned orally by one participant. The discussion on the limitations (cf. Section 6) also touches on the completeness of criteria.

[3]translated from German.

As opposed to the algorithm introduced in [2], a feature-centric ranking algorithm should incorporate information about the queried feature itself and the surrounding environment. Although the queried feature lies within the viewshed of the video, the view may be impaired by other features, such as a building which lies between the camera and the query feature. It is therefore necessary that a feature-centric ranking algorithm considers additional information about the geographic properties of other features in the scene.

Most characteristics listed in Table 1 are directly related to the visibility of the query feature. However, some characteristics focus less on the query feature and its spatial properties but more on general camera work and the properties of the footage. Since this article focuses on a feature-centric approach which relies solely on available metadata, the implementation of "Camera Work", and "Focus/Sharpness" does not fall within the scope of the work. Furthermore, the criteria "Camera Height", "Surroundings", "Perspective (3D)", "Distracting Objects", and "Dimension/Relation to other Objects" may be assessed using contextual information, but their detailed discussion is left for future work. Four criteria were implemented during the work and are further discussed in the remainder of the paper, namely:

- *Feature Depiction* (cf. Section 4.1),
- *Feature Illumination* (cf. Section 4.2)
- *Feature Visibility Duration* (cf. Section 4.3), and
- *Distance to Feature* (cf. Section 4.4),

While three characteristics can be directly inferred from the contextual metadata, the "Feature Illumination" uses an indirect method. A brightness analysis would require analyzing the pixel data (and is computationally expensive). Instead, we estimate the ambient light by evaluating the sun's position throughout the footage. This allows a computationally cheap estimation of the amount of sunlight in the captured scene. From the geographic position and the recording hour, we derive the vertical and horizontal angles of the sun, which in turn indicate the amount of shading in the scene.

The next section provides a detailed description of the four criteria and their implementation in five relevance ranking algorithms.

4 ALGORITHMS FOR FEATURE-CENTRIC VIDEO RANKING

This section describes the technical concepts of our algorithm for ranking georeferenced videos based on a selected feature, its spatial representations, and the videos' geospatial metadata. In Section 3 a focus group has identified 12 characteristics, from which four have been selected for implementation. These four characteristics are translated into five metrics (cf. Table 3). We also implemented three metrics proposed by Ay et al. [2] for comparison and evaluation. We provide the implementation of the algorithms on GitHub[4] as well as demonstration video and further complementary material in our university campus cloud[5].

Our algorithm creates ranking scores for each metrics and FOVScene. A ranking score for a single video comprises the summation and normalization of the individual ranking scores of each scene.

[4]GitHub Repository: https://github.com/sitcomlab/GeoreferencedVideoRanking
[5]Complementary files: https://uni-muenster.sciebo.de/index.php/s/KsZLIqG52fjjak0

Table 1: Relevance criteria identified by the focus group.

Criteria	Description	Score
Feature Depiction	Visibility of the desired features, i.e. whether it is occluded or exceeds the field of view	36
Camera Work	Smoothness of camera movements	31
Focus/Sharpness	Sharpness of features throughout the video	30
Feature Illumination	Brightness of the captured scenes	22
Surroundings	Other features on the footage that support orientation	22
Feature Visibility Duration	Ratio between video duration and visible time of the target feature	20
Perspective (3D)	Angle from which the query feature has been recorded	16
Location (Orientation to query feature)	Combination of perspective and distance	15
Dimension / Relation to other features	Whether the relative feature size with respect to nearby features can be determined	14
Distracting Features	Amount of distraction caused by other features on the footage	13
Distance to Feature	Camera-feature distance	11
Camera Height	Camera height above the ground	10

Table 2: Summary of terms.

Term	Description
B_Q^R	the feature's right-most visible vertex
B_Q^L	the feature's left-most visible vertex
D	the actual feature-camera distance
D^+	the optimal distance
ΔD	the distance deviation
$\vec{d}$	the viewing direction of $V_k^F(t_i)$
$\vec{d} - \theta/2$	angle of left view border
$\vec{d} + \theta/2$	angle of right view border
V_k	a video k
V_k^F	the set of FOVScenes for video k
$V_k^F(t_i)$	a single FOVScene for video k at time i
P	the camera viewpoint
Q	the spatial representation of the queried feature
O_x	a feature x within $V_k^F(t_i)$
Φ	the solar azimuth angle
θ	the horizontal viewing angle
t	the overall video duration
$Q^{\vdash\dashv}$	the actual visible angular range of Q
$Q_{max}^{\vdash\dashv}$	the maximum visible angular range of Q
$O_x^{\vdash\dashv}$	the visible angular range of object O_x (i.e. its occlusion range with respect to Q)

Table 3: Summary of the implemented metrics for the four selected criteria.

Metrics	Description
Distance to Feature	
R_{Dist}	Average distance deviation of a video's *FOVScenes* with respect to D^+
Feature Illumination	
R_{Az}	Average azimuthal difference of a video's *FOVScenes* measured between $\vec{d}$ and Φ
R_{El}	Elevation angle measured for the *FOVScene* at the time half of the video
Feature Depiction	
R_{SA}[6]	Summed overlap area of a video's *FOVScenes* and the query region
R_{TA}[6]	Total overlap area of a video's *FOVScenes* and the query region.
R_{Dep}	Average fraction of the feature of interest that is visible throughout a video
Feature Visibility Duration	
R_{Vis}	Relative visibility duration of a query feature throughout a video
R_D[6]	Overlap duration of a video's *FOVScenes* with the query region

[6] the metrics R_{SA}, R_{TA}, and R_D have been proposed in [2] and were included for comparison and evaluation.

4.1 Feature Depiction Metric R_{Dep}

With R_{SA} and R_{TA}, Ay et al. [2] presented two metrics for evaluating how well a video captures a specific query region. The metrics relied on the estimated overlap region of a particular video and the query region. However, a query feature is different from a query

region. The feature represents a single object such as a building, which may be occluded by other nearby features (i.e. other buildings). To account for such occlusions, we proposed the metric R_{Dep} to evaluate the depiction of the queried feature within a particular scene. R_{Dep} measures the relative portion of Q that is not occluded by other features within a particular $V_k^F(t_i)$. Let Q be the spatial representation of the queried feature and P the camera viewpoint. By determining the viewing angles between P and the border points of Q, the algorithm defines the angular range of Q within $V_k^F(t_i)$, which is denoted as $Q^{\vdash\dashv}$. The maximum angular range is given by the angular range of $V_k^F(t_i)$ itself, $[\vec{d}-\theta/2, \vec{d}+\theta/2]$. This is a limitation, which accounts for cases where Q intersects the view borders and thus overflows the visible scene. $Q^{\vdash\dashv}$ is then reduced by the angular range of every other feature O_x within $V_k^F(t_i)$, $O_x^{\vdash\dashv}$. Since these represent the occluded portions of Θ with respect to Q within $V_k^F(t_i)$, we also refer to them as occlusion ranges. Sub-ranges may originate in cases where the angular range of Q completely contains a particular occlusion range. The scene features to compute R_{Dep} are retrieved dynamically from the OpenStreetMap Overpass API[7]. For simplification, only features of type building are queried. However, any other feature class of geometry type polygon may also be processed. To improve the performance, several filter steps reduce the amount of comparisons between the features' angular ranges and Q. All features intersecting with the current V_k^F are queried from the Overpass API and cached in a feature set O during the computation of the ranking scores. For each $V_k^F(t_i)$ the features intersecting with $V_k^F(t_i)$ are derived from the set. Then, a convex hull polygon is constructed from the vertices of Q and P. Finally, angular ranges relative to $\vec{d}$ are computed for all the remaining features whose geometries intersect with the hull polygon. These in turn are used for computing R_{Dep} for the respective $V_k^F(t_i)$.

4.2 Feature Illumination Metrics R_{Az}, R_{El}

The visibility of a feature throughout a footage is heavily dependent on how well it is illuminated. To determine the feature illumination we propose the two different metrics R_{Az} and R_{El}. Both metrics leverage the sun's topographic position with respect to the camera viewpoint.

$$R_{Az}\left(V_k^F, Q\right) = \sum_{i=1}^{n} \min\left(\left|(\vec{d}_{t_i} - \Phi)\right|, 360 - \left|(\vec{d}_{t_i} - \Phi)\right|\right) \quad (2)$$

Using equation (2), R_{Az} leverages the azimuth angle Φ to compute the horizontal solar position relative to $\vec{d}$. In contrast, R_{El} corresponds to the vertical position of the sun, i.e. the solar elevation angle e. In comparison R_{Az} determines whether a $V_k^F(t_i)$ was recorded in backlight or not while R_{El} estimates the degree to which a particular V_k^F is dominated by shadow.

For computing the topographic solar position a number of solar equations are used, which were presented by Grena [10]. He states that the equations offer a well suited trade-off between computational performance and precision [10]. Since minor errors in the computed solar position do not heavily affect the algorithm scores the precision offered by the equations is sufficient for estimating the illumination setting of a video. The solar position is computed only once per every V_k^F since significant changes only occur very slowly and the average video duration is expected to be much shorter than that. Therefore, R_{El} is only measured at the time half of the video. As opposed to this, R_{Az} is computed for each individual $V_k^F(t_i)$ since $\vec{d}$ may change greatly among a video which in turn affects the relative horizontal solar position of the respective scene.

4.3 Feature Visibility Duration Metric R_{Vis}

Based on the score computed for R_{Dep} (cf. chapter 4.1) the metric R_{Vis} measures the visibility duration of Q throughout a particular V_k^F. It is determined by dividing the period for which $R_{Dep} > 0$ is true by the overall video duration as outlined by equation (3). Note that in the equation, t denotes the overall duration of V_k^F and t_i denotes the duration of a particular $V_k^F(t_i)$.

$$R_{Vis}\left(V_k^F, Q\right) = \sum_{i=1}^{n} \frac{(t_{i+1} - t_i)}{t}, \quad \text{for } i \text{ when } R_{Dep}\left(V_k^F(t_i), Q\right) > 0 \quad (3)$$

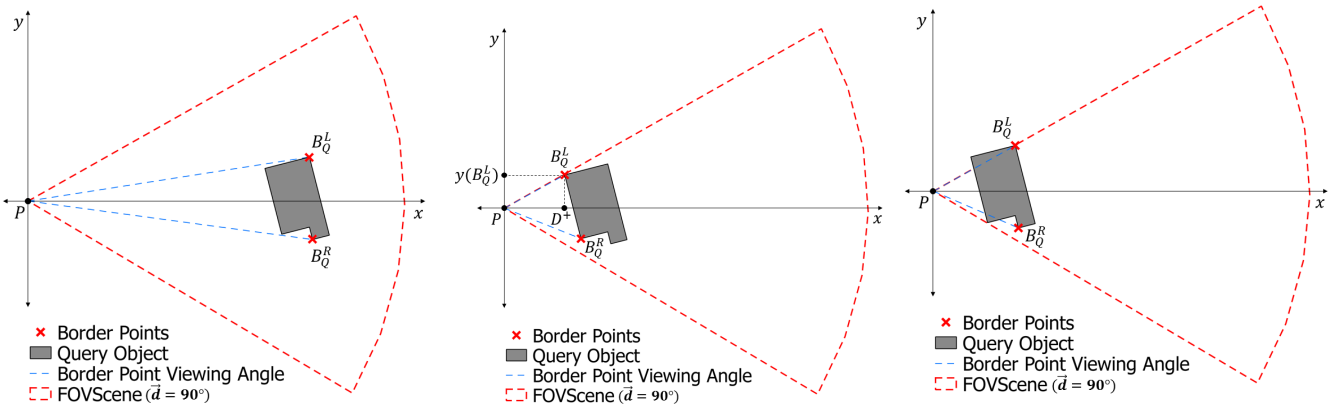
The accuracy of R_{Vis} thereby relies on the sampling rate of V_k^F. Furthermore, R_{Vis} is similar to the metric R_D that was proposed by Ay et al. [2]. However, R_{Vis} relies on the estimated viewing range of Q within a particular $V_k^F(t_i)$ rather than on the spatial overlap between both geometries.

4.4 Distance to Feature Metric R_{Dist}

The distance between the camera viewpoint P and the queried feature is evaluated by determining how well the feature fills a particular video scene. It was conjectured that in a perfect setting the queried feature should be located as close as possible to P without overflowing the field of view. Therefore, the algorithm R_{Dist} estimates the optimal distance D^+ at which the feature's left-most visible vertex B_Q^L and right-most visible vertex B_Q^R, respectively, would intersect with the view borders of $V_k^F(t_i)$ for the first time when moving the feature towards P along the viewing direction $\vec{d}$. Note that due to perspective issues, the border points of Q may be different for different distances between P and Q. In such cases, the exact value of D^+ is not determined correctly. However, the resulting error lies within an acceptable extent so that no major inaccuracies are to be expected. Figure 1c exemplifies this situation. Next, R_{Dist} computes the difference between the actual camera-feature distance and the optimal camera-feature distance D^+. The resulting value is then subtracted from D^+. A scene $V_k^F(t_i)$ with a camera–feature–distance close to D^+ would thus obtain a high ranking score.

Let Q be again the spatial representation of the queried feature and $\vec{d}$ be the viewing direction of $V_k^F(t_i)$ with a horizontal viewing angle θ. This implies that the angles of the left and right view borders are $(\vec{d} - \theta/2)$ and $(\vec{d} + \theta/2)$, respectively. To find the intersections between Q and the view borders a cartesian coordinate system centered at P with $\vec{d} = 90°$ can be used (see Figure 1). The view borders can then be represented as linear functions of the form $y = m * x + b$ with y-axis section $b = 0$ and gradient m given

[7] OSM Overpass API: http://overpass-api.de



(a) Initially computed border points **(b) Actual border points and optimal distance** **(c) Error caused by border point shift**

Figure 1: Calculation of D^+

by $tan(\vec{d} - \theta/2)$ and $tan(\vec{d} + \theta/2)$, respectively. D^+ can now be computed by equating the terms of the view borders with the parallels cutting B_Q^L and B_Q^R with gradient $m = 0$ and y being the respective point's northing value.

Algorithm 1 Complete Depiction, defined as deviation from the optimal distance to the feature. $DistanceDeviation(V_k^F(t_i), Q, B)$

$\alpha_{L_0} = LimitDegrees180((Angle(P, B_Q^L) + (90 - d)))$
$\alpha_{R_0} = LimitDegrees180((Angle(P, B_Q^R) + (90 - d)))$
$L = Dest(Lat_P, Lng_P, Dist(P, B_Q^L), 90 - (\alpha_{B_Q^R} - \alpha_{B_Q^L})/2)$
$R = Dest(Lat_P, Lng_P, Dist(P, B_Q^R), 90 - (\alpha_{B_Q^R} - \alpha_{B_Q^L})/2)$
$L^v = Dist(< Lng_P, Lat_L >, P)$
$R^v = Dist(< Lng_P, Lat_R >, P)$
$L^h = Dist(< Lng_L, Lat_P >, P)$
$R^h = Dist(< Lng_R, Lat_P >, P)$
$A_L = TransformAngle(((d - \theta/2) + (90 - d)) \mod 180)$
$A_R = TransformAngle(((d + \theta/2) + (90 - d)) \mod 180)$
$D_L = |L^v / \tan(A_L)|$
$D_R = |R^v / \tan(A_R)|$
if $|(D_L - L^h)| > |(D_R - R^h)|$ **then**
    $D^+ = D_L$
    $D = L^h$
else
    $D^+ = D_R$
    $D = R^h$
end if
$\Delta D = D^+ - |(D - D^+)|$
return ΔD

The greater one of the computed x-values now corresponds to D^+ and is then subtracted from the absolute difference between itself and the actual camera-feature distance to obtain the distance deviation ΔD. These in turn represents the distance score for $V_k^F(t_i)$. Algorithm 1 outlines the computation of ΔD for a particular V_k^F. The subroutines used by the algorithm are as follows:

- *LimitDegrees180*: Limits the given angle to the range [0, 180].
- *Dest*: Computes the destination for a given base point, distance and angle.
- *Dist*: Computes the distance between two points.
- *TransformAngle*: Transforms a geographical direction into an angle within a Cartesian Coordinate System, i.e. an angle measured with respect to the x-axis instead of geographical north.

Finally, the cumulated distance score for a video V_k^F is given by equation (4).

$$R_{Dist}\left(V_k^F, Q\right) = \sum_{i=0}^{n} \Delta D\left(V_k^F(t_i)\right) \tag{4}$$

4.5 Calculating the final scores

All scores except of R_{El} are computed for each individual $V_k^F(t_i)$ of a given video V_k. To obtain aggregated scores for the whole video, the individual frame scores need to be summed up. To allow for comparison of the relevance scores of multiple videos, their scores further need to be normalized by the number of $V_k^F(t_i)$. Otherwise longer videos would outperform shorter ones since more computed scores are summed up. Therefore, the scores R_{Az} and R_{Dist} are divided by the number of FOVScenes of V_k $|V_k^F(t_i)|$. As opposed to this, R_{Dep} and R_{Vis} already return scores relative to the maximum angular range of Q and the video duration t, respectively. However, since R_{Dep} is only meaningful for FOVScenes where Q can be seen at all it is divided only by the number of $V_k^F(t_i)$ where $R_{Dep} > 0$. Note that due to its temporal nature R_{Vis} is divided by the overall duration of V_k^F instead of the number of frames. For convenience, the resulting scores of R_{Dep} and R_{Vis} are multiplied with 100 to obtain percentage values. The normalized rank scores can then be used for video ranking purposes with respect to the different characteristics.

5 EVALUATION

The evaluation of the metrics from Table 3 is the subject of this section. The evaluation addresses two aspects: cognitive plausibility,

and computational performance. In addition, the comparison of the MAP (Mean Average Precision) scores of the metrics gives some preliminary indications about the similarity of some of the metrics.

5.1 Cognitive plausibility

Kennedy [19] indicated that a cognitive plausible system is either a system which is capable of performing as well as humans do on cognitive tasks or a system which is built on cognitively plausible components. Here, cognitive plausibility refers to the first aspect of the definition (i.e. performing as good as humans do). The Discounted Cumulated Gain (DCG) of all the criteria was computed, using a base log of 2. Contrary to measures such as precision and recall which do not take into account the position of the retrieved document in the result list, the DCG [12, 13] discounts the relevance value of documents ranked further down in the result list. 14 participants (eight female, six male) were asked to run three queries returning seven videos each, and rank the videos returned according to their relevance. Their ranking was then aggregated using the Borda Count method. A Borda Counting was chosen because it rewards consensus and wide approval [27]. The videos were retrieved from the GeoVid API[8]. Each of the videos showed the Marina Bay Sands hotel in Singapore. None of the study participants was familiar with the area depicted in the video. Figure 2 illustrates the results obtained. With the user rankings provided by the study participants as a baseline, R_{Dep} maximizes cognitive plausibility of the algorithm while R_{Dist} minimizes it. The curve for R_{Vis} is between those of R_{Dep} and R_{Dist}. The NDCG (normalized DCG) scores for R_{Dep}, R_{Vis} and R_{Dist} are 0.949, 0.778 and 0.696 respectively. As Table 1 shows, participants identified depiction to be more important than relative visibility duration, and the latter to be more important than distance to feature. Thus, the implementation of these three criteria in the algorithm is sound with respect to participants' preferences. In addition, measuring the lighting could be done with either R_{Az} or R_{El} (NDCG scores of 0.709 and 0.808 respectively). Keeping Table 1 in mind would suggest that R_{El} is more adequate since participants identified lighting characteristics as slightly more relevant than relative visibility duration.

5.2 Computational performance

Figure 3 shows the computational performance of the different metrics. The time on the Y-axis of the figure was obtained by averaging over 50 queries. The great amount of time needed to compute R_{Dep} and R_{Vis} is mainly due to the complex calculations involved in determining the visible portion of the feature of interest in the video. Moreover, the curves of R_{Dep} and R_{Vis} look similar because R_{Vis} relies on results from R_{Dep}. Another observation from Figure 3 is that R_{Az}, R_{El} and R_D are the less greedy in terms of computational resources. Since R_{Dep} had the highest NDCG score, the metric which offers the best plausibility/performance tradeoff minimizes loss of NDCG while at the same time maximizing gain in time. Let O (for optimum) be that metric, $\alpha = \frac{R_{Dep}}{O}$ the NDCG loss, and $\beta = \frac{T_{RDep}}{T_O}$ the gain in time (T_{RDep} is the time needed to compute R_{Dep}, and T_O is the time needed to compute O). O can be defined as the metric for which $\frac{\alpha}{\beta}$ is minimum. Table 4 shows that R_D is the optimum metric, followed by R_{El}. That is, the overlap duration of the video's FOVScenes with the query region seems to provide the best tradeoff between plausibility and performance. All NDCG values, and execution time values used to compute the plausibility/performance tradeoffs are available at https://uni-muenster.sciebo.de/index.php/s/KsZLIqG52fjjak0.

[8]http://geovid.org/ (last accessed: May 31, 2017).

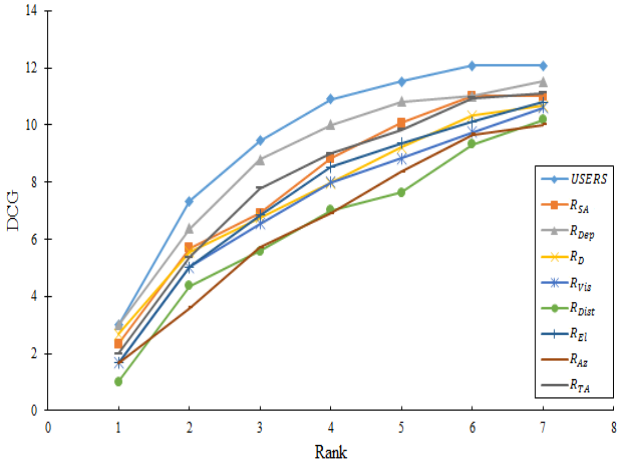


Figure 2: DCG scores of the different ranking metrics.

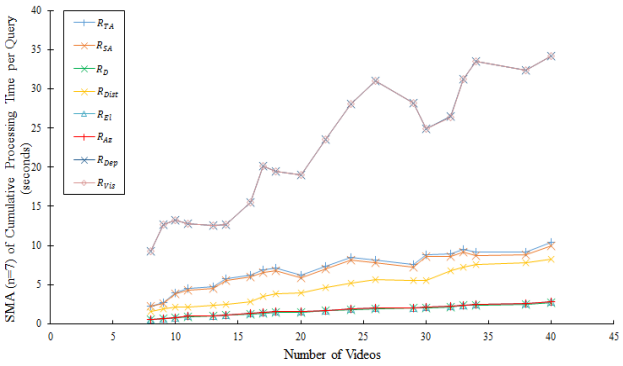


Figure 3: Algorithm performance for different numbers of videos.

Table 4: PaL and PeG ratios with respect to R_{Dep}. PaL refers to the plausibility loss, and PeG to the performance gain.

	$\frac{R_{Dep}}{R_D}$	$\frac{R_{Dep}}{R_{Vis}}$	$\frac{R_{Dep}}{R_{Az}}$	$\frac{R_{Dep}}{R_{El}}$	$\frac{R_{Dep}}{R_{Dist}}$	$\frac{R_{Dep}}{R_{TA}}$	$\frac{R_{Dep}}{R_{SA}}$
PaL	1.16	1.22	1.34	1.17	1.36	1.10	1.10
PeG	14.22	1.00	13.63	13.64	13.47	3.26	3.39
$\frac{PaL}{PeG}$	0.08	1.22	0.10	0.09	0.10	0.34	0.32

5.3 Similarity between metrics

As mentioned at the beginning of the article, the goal of this work is to take advantage of sensor metadata to provide a feature-centric

Table 5: MAP scores for selected pairs of ranking metrics

	N=1	N=2	N=3	N=4	N=5	N=6	N=7
R_{Dep} and R_{Vis}	0	0.333	0.444	0.75	0.733	0.889	1
R_{Vis} and R_{Az}	0	0.333	0.778	0.667	0.733	0.889	1
R_{SA} and R_{Dep}	0.333	0.5	0.556	0.833	0.867	0.889	1
R_D and R_{Vis}	0.333	0.667	0.667	1	0.933	0.889	1
R_{TA} and R_{Dep}	0	0.333	0.667	0.75	0.867	0.889	1

retrieval of videos. The different metrics suggested rely on different types of metadata and yield different outcomes. This section briefly explores the question whether some of these metrics return sufficiently similar results to be substituted, should the necessity arise. This question is pertinent because some metadata may not be available at all, or the available metadata may be erroneous. In this situation, using a metric as a proxy for another one can provide not entirely accurate, but still useful information. To provide a preliminary answer to this question, the MAP (Mean Average Precision) scores were computed for some selected pairs of metrics. The MAP scores indicate the amount of identical videos among several ranking lists, for different values of N. The results obtained are summarized in Table 5. The table shows some similarity between the results returned by R_D introduced in [2], and R_{Vis} proposed in this work. Likewise, both R_{SA} from [2], and R_{Dep} have at least half of the videos within the top N results in common for any N >= 2. These observations are preliminary, but also promising and could be further investigated in future work.

6 LIMITATIONS

The user study evaluating the cognitive plausibility of the algorithm contained seven videos because of a necessary trade-off between the complexity of the ranking tasks, and the duration of the user study sessions. Each session lasted about 60 minutes. Including more videos in the user study does not necessarily provide better insight because several participants confessed being exhausted after viewing all the videos and creating multiple ranking orders. It is challenging to isolate the actual effects of tiredness on the user rankings. Follow-up studies could mitigate these effects by (a) shortening the video ranking sessions, (b) extending the sessions with more videos while including breaks for the participants, or (c) recruit a much larger group of participants, and distribute ranking tasks across participants with similar backgrounds.

In addition, the algorithms use metadata (i.e. video location and orientation) as input. Its effectiveness depends therefore on the accuracy of these metadata. Location accuracy can be improved by techniques such as smoothing splines (for an example, see [47]). Ongoing work [44, 47] is looking at methods to improve video orientation accuracy. Modeling the impact of metadata accuracy on the ranking performance of the algorithms is an interesting issue for future work. Finally, the criteria for feature-centric video ranking were suggested by a relatively small group of participants, with a fairly homogeneous cultural background (i.e. all participants were German). Conducting additional focus groups with more participants, from additional age groups (e.g. 27+), and cultural backgrounds (e.g. Asia, America, Africa) will shed light on criteria which are peculiar to the focus group of this work, and those which have a more universal applicability.

7 CONCLUSION

This article proposed five ranking algorithms to query georeferenced videos for a specific feature based on the videos' spatio-temporal metadata. We conducted a focus group with 6 participants, in which we compiled 12 relevance criteria for feature-centric video ranking. Four of these criteria have been selected and were implemented in five ranking algorithms. We evaluated the algorithms regarding their computational efficiency and user perceived plausibility. The evaluation suggests that the "Feature Visibility Duration" of the video's viewshed with the queried feature geometry offers a good trade-off between computationally performant and cognitive plausible ranking.

The relevance criteria identified by the focus group can be used to make research on feature-centric video search more user-centric. Directions for future work include extending the technical implementation to include all criteria identified by the focus group and conducting additional focus group studies to get a greater understanding of users' wishes as regards feature-centric video search.

ACKNOWLEDGMENTS

Auriol Degbelo gratefully acknowledges funding from the European Union through GEO-C (H2020-MSCA-ITN-2014, 642332, http://www.geo-c.eu/).

REFERENCES

[1] Sakire Arslan Ay, Lingyan Zhang, Seon Ho Kim, Ma He, and Roger Zimmermann. 2009. GRVS: a georeferenced video search engine. In *Proceedings of the seventeen ACM international conference on Multimedia - MM '09*, Wen Gao, Yong Rui, Alan Hanjalic, Changsheng Xu, Eckehard G. Steinbach, Abdulmotaleb El-Saddik, and Michelle X. Zhou (Eds.). ACM Press, Beijing, China, 977.

[2] Arslan Ay, Roger Zimmermann, and Seon Ho Kim. 2010. Relevance Ranking in Georeferenced Video Search. *Multimedia Syst.* 16, 2 (March 2010), 105–125. https://doi.org/10.1007/s00530-009-0177-x

[3] S Ay, L Zhang, S Kim, M He, and R Zimmermann. 2009. GRVS: A georeferenced video search engine. In *Proceedings of the 17th ACM international conference on Multimedia*. New York, NY, USA, 977–978.

[4] Sakire Arslan Ay, Roger Zimmermann, and Seon Ho Kim. 2008. Viewable scene modeling for geospatial video search. In *Proceeding of the 16th ACM international conference on Multimedia - MM '08*. ACM, New York, NY, USA, 309–318.

[5] Rosanna L. Breen. 2006. A Practical Guide to Focus-Group Research. *Journal of Geography in Higher Education* 30, 3 (Nov. 2006), 463–475. https://doi.org/10.1080/03098260600927575

[6] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image retrieval. *Comput. Surveys* 40, 2 (2008), 1–60.

[7] Tobias Emrich, Olivia Hofer, Andreas Kolb, Johannes Niedermayer, Nepumuk Seiler, and Michael Weiler. 2015. Video route. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15*, Jie Bao, Christian Sengstock, Mohammed Eunus Ali, Yan Huang, Michael Gertz, Matthias Renz, and Jagan Sankaranarayanan (Eds.). ACM Press, Bellevue, Washington, USA, 1–4.

[8] Flora Gilboa-Solomon, Gal Ashour, and Ophir Azulai. 2013. Efficient storage and retrieval of geo-referenced video from moving sensors. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'13*, Craig A. Knoblock, Markus Schneider, Peer Kröger, John Krumm, and Peter Widmayer (Eds.). ACM Press, Orlando, Florida, USA, 394–397.

[9] Carlos Granell and Frank O. Ostermann. 2016. Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. *Computers, Environment and Urban Systems* 59 (Sept. 2016), 231–243.

[10] Roberto Grena. 2008. An algorithm for the computation of the solar position. *Solar Energy* 82, 5 (2008), 462–470.

[11] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. 2011. A Survey on Visual Content-Based Video Indexing and Retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 6 (2011), 797–819.

[12] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[13] Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *Proceedings of the 30th European Conference on Information Retrieval*, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R.W. White (Eds.). Springer Berlin Heidelberg, Glasgow, UK, 4–15.

[14] Henry Jenkins, Joshua Green, John Hartley, and Jean Burgess. 2013. *Youtube: online video and participatory culture*. Polity Press, Place of publication not identified. http://public.eblib.com/choice/publicfullrecord.aspx?p=4029558

[15] Rongrong Ji, Yue Gao, Wei Liu, Xing Xie, Qi Tian, and Xuelong Li. 2015. When Location Meets Social Multimedia: A Survey on Vision-Based Recognition and Mining for Geo-Social Multimedia Analytics. *ACM Trans. Intell. Syst. Technol.* 6, 1 (March 2015), 1:1–1:18.

[16] Simon Jones and Ling Shao. 2013. Content-based retrieval of human actions from realistic video databases. *Information Sciences* 236 (2013), 56–65.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*. 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462

[18] Pascal Kelm, Sebastian Schmiedeke, and Lutz Goldmann. 2015. Imcube @ MediaEval 2015 Placing Task: A Hierarchical Approach for Geo-referencing Large-Scale Datasets. In *Working Notes Proceedings of the MediaEval 2015 Workshop*. Martha Larson et al, Wurzen, Germany, 3.

[19] William G Kennedy. 2009. "Cognitive plausibility" in cognitive modeling, artificial intelligence, and social simulation. *Proceedings of the International Conference on Cognitive Modeling - ICCM 2009* (2009), 454–455.

[20] Joon-Seok Kim, Seon Ho Kim, and Ki-Joune Li. 2013. Automatic geotagging and querying of indoor videos. In *Proceedings of the Fifth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness - ISA '13*, Craig A. Knoblock, Markus Schneider, Peer Kröger, John Krumm, and Peter Widmayer (Eds.). ACM Press, Orlando, Florida, USA, 50–53.

[21] Kyong-ho Kim, Sung-soo Kim, Sung-ho Lee, Jong-hyun Park, and Jong-hun Lee. 2003. The Interactive Geographic Video. In *IEEE International Geoscience and Remote Sensing Symposium. Proceedings*, Vol. 1. IEEE, New York, NY, USA, 59–61.

[22] Youngwoo Kim, Jinha Kim, and Hwanjo Yu. 2012. GeoSearch : Georeferenced Video Retrieval System. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 1540–1543.

[23] Youngwoo Kim, Jinha Kim, and Hwanjo Yu. 2014. GeoTree: Using spatial information for georeferenced video search. *Knowledge-Based Systems* 61, May 2014 (2014), 1–12.

[24] Youngwoo Kim, Jinha Kim, and Hwanjo Yu. 2014. GeoTree: Using spatial information for georeferenced video search. *Knowledge-Based Systems* 61 (May 2014), 1–12. http://www.sciencedirect.com/science/article/pii/S0950705114000549

[25] Richard A. Krueger and Mary Anne Casey. 2015. *Focus groups: a practical guide for applied research* (5th edition ed.). SAGE, Thousand Oaks, California.

[26] Dongha Lee, Jinoh Oh, Woong-Kee Loh, and Hwanjo Yu. 2016. GeoVideoIndex: Indexing for georeferenced videos. *Information Sciences* 374 (Dec. 2016), 210–223.

[27] Jonathan Levin and Barry Nalebuff. 1995. An introduction to vote-counting schemes. *The Journal of Economic Perspectives* 9, 1 (1995), 3–26.

[28] Paul Lewis, Stewart Fotheringham, and Adam Winstanley. 2011. Spatial video and GIS. *IJGIS* 25, 5 (2011), 697–716.

[29] Lin Tzy Li, Javier A.V. Muñoz, Jurandy Almeida, Rodrigo T Calumby, Otávio AB Penatti, Ícaro C Dourado, Keiller Nogueira, Pedro R Mendes-Junior, Luís AM Pereira, Daniel CG Pedronette, Jefersson A. dos Santos, Marcos A. Gonçalves, and Ricardo da S. Torres. 2015. RECOD@ Placing Task of MediaEval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop (CEUR-WS)*. Martha Larson et al, Wurzen, Germany, 3. http://ceur-ws.org/Vol-1436/

[30] Lin Tzy Li, Daniel Carlos Guimarães Pedronette, Jurandy Almeida, Otávio A. B. Penatti, Rodrigo Tripodi Calumby, and Ricardo da Silva Torres. 2014. A rank aggregation framework for video multimodal geocoding. *Multimedia Tools and Applications* 73, 3 (Dec. 2014), 1323–1359. https://doi.org/10.1007/s11042-013-1588-4

[31] Ying Lu, Cyrus Shahabi, and Seon Ho Kim. 2014. An efficient index structure for large-scale geo-tagged video databases. In *Proc. SIGSPATIAL '14*, Yan Huang, Markus Schneider, Michael Gertz, John Krumm, and Jagan Sankaranarayanan (Eds.). ACM Press, Dallas, Texas, USA, 465–468.

[32] Ying Lu, Hien To, Abdullah Alfarrarjeh, Seon Ho Kim, Yifang Yin, Roger Zimmermann, and Cyrus Shahabi. 2016. GeoUGV: user-generated mobile video dataset with fine granularity spatial metadata. In *Proceedings of the 7th International Conference on Multimedia Systems - MMSys '16*, Christian Timmerer (Ed.). ACM Press, Klagenfurt, Austria, 1–6.

[33] Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew Gallagher. 2011. Geotagging in multimedia and computer vision-a survey. *Multimedia Tools and Applications* 51, 1 (2011), 187–211.

[34] R. Madison and Yuetian Xu. 2010. Tactical geospatial intelligence from full motion video. In *Applied Imagery Pattern Recognition Workshop (AIPR), 2010 IEEE 39th*. 1–6. https://doi.org/10.1109/AIPR.2010.5759699

[35] F. O. Ostermann, H. Huang, G. Andrienko, N. Andrienko, C. Capineri, Károly Farkas, and R. S. Purves. 2015. Extracting and Comparing Places Using Geo-social Media. In *Proc. ISSDQ 2015*. La Grande Motte, France, 1–6.

[36] J. Ross and C. Coman. 2014. Full Motion Video in a coalition environment. In *2014 10th International Conference on Communications (COMM)*. 1–4. https://doi.org/10.1109/ICComm.2014.6866709

[37] Beomjoo Seo, Jia Hao, and Guanfeng Wang. 2011. Sensor-rich video exploration on a map interface. In *Proceedings of the 19th ACM international conference on Multimedia - MM '11*. ACM, New York, NY, USA, 791–792.

[38] Kim Shearer, Svetha Venkatesh, and Dorota Kieronska. 1996. Spatial Indexing for Video Databases. *Journal of Visual Communication and Image Representation* 7, 4 (1996), 325–335. https://doi.org/10.1006/jvci.1996.0028

[39] Zhijie Shen, Sakire Arslan Ay, Seon Ho Kim, and Roger Zimmermann. 2011. Automatic Tag Generation and Ranking for Sensor-rich Outdoor Videos. In *Proceedings of the 19th ACM international conference on Multimedia (MM '11)*. ACM, New York, NY, USA, 93–102.

[40] Sarah L. Smiley, Andrew Curtis, and Joseph P. Kiwango. 2017. Using Spatial Video to Analyze and Map the Water-Fetching Path in Challenging Environments: A Case Study of Dar es Salaam, Tanzania. *Tropical Medicine and Infectious Disease* 2, 2 (April 2017), 8. https://doi.org/10.3390/tropicalmed2020008

[41] Benjamin W. Tatler. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* 7, 14 (Nov. 2007), 4–4. https://doi.org/10.1167/7.14.4

[42] M. van Persie, A. Oostdijk, J. Fix, M. C. van Sijl, and L. Edgardh. 2012. Real-Time UAV based Geospatial Video integrated into the Fire Brigades Crisis Management GIS System. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XXXVIII-1/C22 (Sept. 2012), 173–175.

[43] Max A Viergever, Remco C Veltkamp, and Hans Burkhardt. 2013. *State-of-the-Art in Content-Based Image and Video Retrieval*. Springer Science & Business Media.

[44] Guanfeng Wang, Yifang Yin, Beomjoo Seo, Roger Zimmermann, and Zhijie Shen. 2013. Orientation data correction with georeferenced mobile videos. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'13*, Craig A. Knoblock, Markus Schneider, Peer Kröger, John Krumm, and Peter Widmayer (Eds.). ACM Press, Orlando, Florida, USA, 400–403.

[45] Yifang Yin, Beomjoo Seo, and Roger Zimmermann. 2015. Content vs. Context: Visual and Geographic Information Use in Video Landmark Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11, 3 (2015), 39:1–39:21. https://doi.org/10.1145/2700287

[46] Yifang Yin, Zhijie Shen, Luming Zhang, and Roger Zimmermann. 2015. Spatial-Temporal Tag Mining for Automatic Geospatial Video Annotation. *ACM Trans. Multimedia Comput. Commun. Appl.* 11, 2 (Jan. 2015), 29:1–29:21.

[47] Yifang Yin, Guanfeng Wang, and Roger Zimmermann. 2016. Automatic geographic metadata correction for sensor-rich video sequences. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '16*, Siva Ravada, Mohammed Eunus Ali, Shawn D. Newsam, Matthias Renz, and Goce Trajcevski (Eds.). ACM Press, Burlingame, California, USA, 1–10.

[48] Yifang Yin, Yi Yu, and Roger Zimmermann. 2015. On Generating Content-Oriented Geo Features for Sensor-Rich Outdoor Video Search. *IEEE Transactions on Multimedia* 17, 10 (2015), 1760–1772.

[49] Bo Zhang, Qinlin Li, Hongyang Chao, Bill Chen, Eyal Ofek, and Ying-Qing Xu. 2010. Annotating and Navigating Tourist Videos. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10)*. ACM, New York, NY, USA, 260–269.

[50] Ying Zhang and Roger Zimmermann. 2012. DVS: A Dynamic Multi-video Summarization System of Sensor-rich Videos in Geo-space. In *Proc. 20th Conf. on Multimedia (MM '12)*. ACM, New York, NY, USA, 1317–1318.