Proc. OASIS Workshop @ 32nd ACM Conf. on Hypertext and Social Media (ACM HT'21), Aug.-Sept. 2021. Authors' accepted version.

Automatic and Semi-automatic Augmentation of Migration Related Semantic Concepts for Visual Media Retrieval

Damianos Galanopoulos
dgalanop@iti.gr
CERTH-ITI
Thermi, Greece

Erick Elejalde
elejalde@l3s.de
L3S, Leibniz-University
Hannover, Germany

Alexandros Pournaras
apournaras@iti.gr
CERTH-ITI
Thermi, Greece

Claudia Niederée
niederee@l3s.de
L3S, Leibniz-University
Hannover, Germany

Vasileios Mezaris
bmezaris@iti.gr
CERTH-ITI
Thermi, Greece

ABSTRACT

Understanding the factors related to migration, such as perceptions about routes and target countries, is critical for border agencies and society altogether. A systematic analysis of communication and news channels, such as social media, can improve our understanding of such factors. Videos and images play a critical role in social media as they have significant impact on perception manipulation and misinformation campaigns. However, more research is needed in the identification of semantically relevant visual content for specific queried concepts. Furthermore, an important problem to overcome in this area is the lack of annotated datasets that could be used to create and test accurate models. A recent study proposed a novel video representation and retrieval approach that effectively bridges the gap between a substantiated domain understanding - encapsulated into textual descriptions of Migration Related Semantic Concepts (MRSCs) - and the expression of such concepts in a video. In this work, we build on this approach and propose an improved procedure for the crucial step of the concept labels' textual augmentation, which contributes towards the full automation of the pipeline. We assemble the first, to the best of our knowledge, migration-related videos and images dataset and we experimentally assess our method on it.

CCS CONCEPTS

• **Information systems → Learning to rank**; **Multimedia and multimodal retrieval**; **Information retrieval query processing**.

KEYWORDS

Semantic queries; Migration Related Semantic Concepts; Image/video retrieval

ACM Reference Format:
Damianos Galanopoulos, Erick Elejalde, Alexandros Pournaras, Claudia Niederée, and Vasileios Mezaris. 2021. Automatic and Semi-automatic Augmentation of Migration Related Semantic Concepts for Visual Media Retrieval. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks (OASIS '21), August 30, 2021, Virtual Event, Ireland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3472720.3483618

1 INTRODUCTION

Migration is a very complex and critical issue around the globe. Causes that bring people to migrate are diverse, ranging from war, political instability, or economic depression to environmental reasons and job opportunities. The perception of possible migration routes and opportunities in target countries plays an essential role in migration-related decision-making. With the effective use of multimodal elements (e.g., text, images, and videos), social media might impact these decisions, especially when it comes to irregular migration. These media channels may manipulate perceptions and, often, lead to misperceptions (e.g., targeted misinformation campaigns). This creates the need among policy-makers and border control agencies to better *understand migration-decision factors and how they are expressed* in different modalities. Given the overwhelming amount of content, automated analysis and evaluation of the available media from various sources are necessary to predict and prevent possible misperception-related risks that affect migrants.

However, identifying how Migration Related Semantic Concepts (MRSCs) are expressed in the social sphere using visual media (i.e., videos or images) is challenging, even for human experts. Building on top of a recently proposed method [5], which is the first and only work so far dealing with migration-related visual media retrieval, we extend the state of the art in two main aspects: i) We design two new strategies for MRSCs augmentation that generate complex sentences to describe them. This augmentation is beneficial for creating a more meaningful input that improves MRSC-based image/video retrieval. ii) We collect migration-related images and videos to compose a MRSC-related dataset. Given the lack of other migration domain-specific datasets, our collection allows us to evaluate the performance of our method in the targeted context.

2 MRSCS DEFINITION AND CLASSIFICATION

Semantic concepts are meaningful entities that form in the mind of a person as a combination of the information received and our unique

OASIS '21, August 30, 2021, Virtual Event, Ireland

ACM ISBN 978-1-4503-8632-6/21/08…$15.00
https://doi.org/10.1145/3472720.3483618

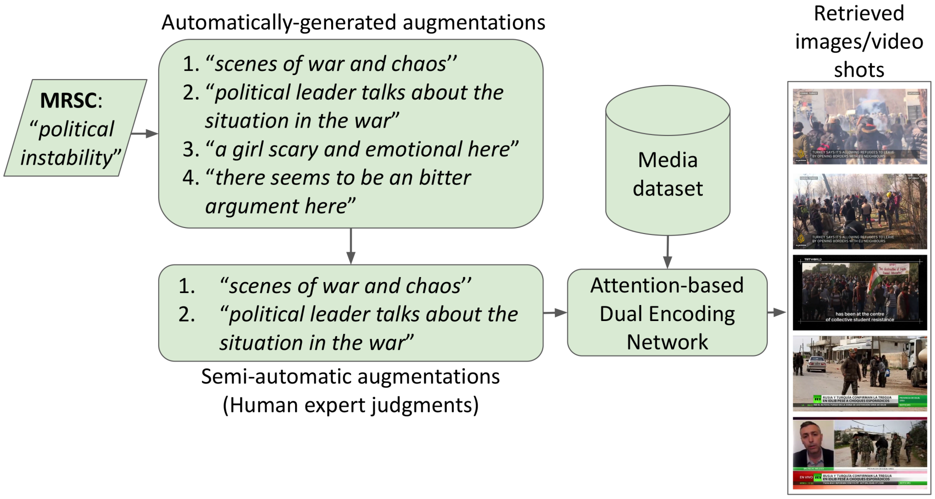
Automatically-generated augmentations
MRSC: "political instability"
1. "scenes of war and chaos''
2. "political leader talks about the situation in the war"
3. "a girl scary and emotional here"
4. "there seems to be an bitter argument here"
1. "scenes of war and chaos''
2. "political leader talks about the situation in the war"
Semi-automatic augmentations (Human expert judgments)
Media dataset
Attention-based Dual Encoding Network
Retrieved images/video shots

Figure 1: An overview of the proposed MRSC retrieval method.

background. More importantly, semantic concepts are intrinsically coupled to a context [6]. For example, "Education" can convey the notion of the active process of learning or the cultural capital of a migrant, which allows her/him to succeed in the host country. Here we focus on Migration Related Semantic Concepts (MRSCs), i.e., concepts relevant in the context of migration.

For our analysis, we use a set of MRSCs based on migration theories and additional input from domain experts, as discussed in [5]. These MRSCs serve as a link between the language used by migration specialists and the identification process of relevant visual information from social media to be interpreted and processed. Moreover, by building over theoretically-founded categories and ideas, this methodology forwards the wide variety of aspects that need to be considered to produce a more complete/unbiased representation of the topics.

The combination of concepts further supports composing multiple aspects and specifies further semantics. This can also be used to define new notions (e.g., see the World bank theme taxonomy[1]). For instance, the aspects "family" and "war" can be combined in a template as "Families in war". This hierarchical pattern is used to organize the MRSCs. Also, it simplifies annotating information at different aggregation levels and the definition of each concept's context.

The MRSCs are grouped into five categories: economic, social, demographic, environmental, and political. These categories act as the upper level of the hierarchy. Below them, as detailed in [5], there are two additional levels: out of a total of 106 MRSCs, 20 are located on the first level and the other 86 directly under them (Table 1 shows some examples). Note that these MRSCs do not constitute a comprehensive list of all the concepts related to migration. Also, this tree-style hierarchical organization does not necessarily make the MRSCs mutually exclusive.

3 MRSCS-BASED VIDEO RETRIEVAL

To identify how the MRSCs are expressed on visual media (i.e., videos and images), we adopted an MRSCs-based video retrieval method as described in [5]. This work addresses the MRSC detection problem as a particular type of Ad-hoc video search (AVS). The goal of AVS is to retrieve unlabeled video shots when a complex textual sentence is used as an input query. The AVS problem's main characteristic is the absence of positive video exemplars for the input query, which could be used for training a supervised classifier. Instead, an AVS method must be capable of handling textual queries in the wild. Since it is not easy to find large amounts of training video exemplars for MRSCs, the AVS method perfectly fits to adopt it as our base system.

3.1 MRSCs Augmentation

As the MRSCs are typically high-level abstractions of semantic concepts, the video-MRSCs correlation is challenging, even for human experts. For example, a human annotator would probably have difficulty deciding which videos should be annotated with the MRSC "Ethnicity". What should be shown in the video to make it relevant to "Ethnicity"? So a procedure that will enrich the MRSC labels with sentences that are explanatory and complementary to the original label, would be beneficial for more efficient video retrieval. In [5], this augmentation procedure was dealt with by employing human experts, who manually augmented a few MRSCs with a set of short but relatively complex sentences to describe them. This process is, however, time-consuming and does not scale to hundreds of potential concepts or dynamically-added new concepts. The present work addresses the MRSCs augmentation problem by proposing two approaches: an automatic and a semi-automatic one. In the automatic approach, for every MRSC, a set of related sentences is identified from a pre-defined large-scale pool of available sentences. Specifically, the relation between an MRSC and the available sentences is calculated as follows: we used the Sentence-BERT method presented in [12] to generate a vector representation for every

[1] https://pubdocs.worldbank.org/en/275841490966525495/Theme-Taxonomy-and-definitions.pdf

Table 1: MRSCs organized in five general categories and in two hierarchical levels (partial listing).

Category	**1st Level**	**2nd Level (examples)**
Economic	Labour market	Working conditions; Labor movements; Job segmentation
	Migrant groups infrastructure	Migration industries; Family labor
	Capital flows	Informal economic activities; International trade
Social	Ethnic minority formation	Others-definition; Self-definition; Ethnic identity
	Cultural capital	Adaptability; Education; Knowledge of other country
	Ethnicity	Language; Culture; Xenophobia; Racism
Demographic	Target-earners	Remittances; Relative success/failure in target country
	Gender	Marriage; Domestic service; Caretaking
	Refugees	War; Political instability; Persecution
Environmental	Urbanization	Access to medical care; Global cities; Stopgaps
	Ecology	Climate change; Pollution; Natural disasters
Political	Settlement	Citizenship; Laws; Nation
	Immigration policies	Representation of immigrants; Change of policies over time
	Crime	Acculturation issues, Ethnic tensions, Cultural predispositions
	Organized crime	Human trafficking; Document fraud, Money laundering

Table 2: Automatically and semi-automatically generated augmentations for the MRSC "Labour market". The italicized (bottom two) augmentations were deleted by the human expert.

MRSC	"Labour market"
Augmentations	workers in a factory
	a group of workers are working in a factory
	a workplace environment people doing work
	men and women working in factory
	people working and sorting
	people are working in a factory
	workers doing their job
	there are some people working in a factory
	in a cloth merchant shop many female employees are working
	a person illustrating how to make people work

sentence. Similarly, a vector representation for each MRSC is also generated, and the *cosine similarity* between every MRSC-sentence pair is calculated. Finally, a list of the k most related sentences is created for every MRSC. In the semi-automatic approach, starting from the above-mentioned automatically-discovered ranked k sentences, human expert judges evaluate and modify the list. To minimize the manual intervention, the experts are allowed to delete sentences from the list but not to add new sentences or modify the existing ones.

An example of the described procedure is presented in Table 2, where the automatically and semi-automatically generated augmentations for the MRSC "Labour market" are shown. As a pool of available sentences, we utilized the video captions of the MSR-VTT dataset [13].

3.2 MRSC-based Video Retrieval

We use a state-of-the-art approach for the AVS problem [7] adapted for MRSC-based video retrieval [5]. This method's overall idea is to train an attention-based dual encoding deep neural network that directly transforms visual and textual content into a common feature space, in which a straightforward comparison between video and textual representations is feasible. The entire network is trained using video-caption pairs and then used as a video retrieval system by inputting MRSCs to recover the most related video shots. The network utilizes two similar modules. Each one consists of multiple encoding levels for the visual and textual content respectively. Additionally, a text-based attention component is used for more efficient textual representation. At training time (using non-migration related datasets, as in [7]) the network translates a media item (an entire video or a video shot) and a textual item (a video caption or a text query) into the new shared feature space, resulting in two new representations, one for the media item and one for the textual. In this space, they are directly comparable. Following the network's training, the MRSCs (with their descriptions) along with images and video shots from the target dataset are used as input to our system. They are encoded into the common feature space, and then for every MRSC, a ranked list with the most closely-related media items within the given image/video dataset can be retrieved. An overview of the proposed method is illustrated in Figure 1.

Table 3: Results of video shot retrieval on the SIN’15 dataset for 30 visual concepts, in terms of XinfAP. The “Concept name” column presents the results when we use as textual input only the concept label. The “Concept name + Manual Augm.” column stands for the setup in which the concept label is manually augmented by a human expert, as in [5]. “Concept name + Aut. Augm.” denotes the proposed automatic procedure for concept augmentation, while the “Concept name + Semi-Aut. Augm.” denotes the proposed semi-automatic method.

SIN’15 dataset				
	Concept name	Concept name + Manual Augm. [5]	Concept name + Aut. Augm.	Concept name + Semi-Aut. Augm.
Airplane	0.3254	**0.5055**	0.4866	0.4952
Anchorperson	0.0067	0.0145	**0.0589**	**0.0589**
Basketball	0.0134	**0.1814**	0.1750	0.1750
Bicycling	0.0569	**0.3730**	0.3317	0.3317
Boat Ship	0.4804	0.5998	**0.6102**	**0.6102**
Bridges	0.0850	**0.1615**	0.1142	0.1154
Bus	0.1215	**0.1382**	0.1133	0.1133
Car Racing	0.0000	**0.0647**	0.0512	0.0512
Cheering	0.0004	0.0687	**0.0907**	**0.0907**
Computers	0.1480	**0.3620**	0.1312	0.2850
Dancing	0.0002	**0.1239**	0.1013	0.1013
Demonstration Or Protest	0.0000	0.2574	**0.3582**	**0.3582**
Explosion Fire	0.1040	**0.1739**	0.1450	0.1450
Government Leader	0.0003	0.1677	**0.1974**	**0.1974**
Instrumental Musician	0.0002	**0.3458**	0.2715	0.2715
Kitchen	0.0805	**0.3400**	0.3245	0.3245
Motorcycle	0.1303	**0.2360**	0.2187	0.2187
Office	0.0546	0.2425	0.2857	**0.2979**
Old People	0.0473	**0.1993**	0.1252	0.1252
Press Conference	0.0001	0.0219	**0.0439**	**0.0439**
Running	0.0008	0.0178	**0.0391**	**0.0391**
Telephones	0.0000	0.3088	**0.3492**	**0.3492**
Throwing	0.0001	0.0485	**0.0698**	**0.0698**
Flags	0.0685	**0.1560**	0.1511	0.1511
Hill	0.0319	0.0675	**0.1305**	**0.1305**
Lakes	0.0577	**0.2033**	0.1807	0.1807
Quadruped	0.0017	**0.2311**	0.0000	0.0124
Soldiers	0.2436	0.3709	**0.4052**	**0.4052**
Studio With Anchorperson	0.0021	0.0393	**0.0919**	**0.0919**
Traffic	0.1372	**0.2046**	0.1624	0.1624
Mean XinfAP	0.0733	**0.2075**	0.1938	0.2001

3.3 MIGRATION-VISUAL Dataset

At present, no domain-specific datasets are publicly available for MRSC-based video retrieval. To evaluate our method specifically with migration-domain content, we created a migration-related videos and images dataset. This MIGRATION-VISUAL dataset was collected through the Media Mining System (MMS) [2].

The dataset includes images and videos from many traditional online sources, such as newspapers, magazines, journals, think-tanks, scientific and policy-making institutions, as well as social media platforms. The dataset was collected to cover the events during which migrants tried to cross the Greek-Turkish border in February and March 2020. The original collection contained a considerable amount of noise. Most of the images were scraped from web pages, and many were unrelated to the actual story (e.g., banners and logos). To remove some of the noise, we kept only jpg images larger than 50 × 50 pixels. Finally, a set of 161.165 images was obtained: 110.209 still images and 50.956 keyframes from 19.232 video shots coming from 482 videos. For evaluating our MRSC-based retrieval method’s results on this dataset, an annotation procedure similar to the one used for ground-truth generation in the TRECVID AVS task [1] was used. Specifically, after our method retrieves a ranked list of 1000 video shots or images for each MRSC, for each list a human expert annotates 100% of the first 200 retrieved images or shots and 20% of the remaining (randomly selected). The human annotator has three different options: positive (the image is relevant to the specified MRSC), negative (the image is not relevant to this MRSC), skipped (in case the annotator is uncertain). If multiple lists are annotated for the same MRSC (created from different model configurations), the aggregated annotation of these lists is used as ground truth for the evaluation.

Table 4: Results of video shot retrieval on the MIGRATION-VISUAL dataset for 10 MRSCs in terms of XinfAP.

MIGRATION-VISUAL dataset			
MRSC	Concept name	Concept name + Aut. Augm.	Concept name + Semi-Aut. Augm.
Urbanisation	0.2368	**0.6135**	0.6096
Laws	0.0803	0.0931	**0.5453**
Crime	0.1102	0.5541	**0.5584**
War	0.4101	0.5267	**0.5268**
Ethnic identity	0.0549	0.3414	**0.4112**
Life	0.1332	0.3291	**0.8221**
Culture	0.0231	0.2406	**0.2820**
Education	0.0193	0.5436	**0.6075**
Politics	0.3991	0.4657	**0.4732**
Work	0.0867	**0.4561**	**0.4561**
Mean XinfAP	0.1554	0.4164	**0.5292**

Table 5: Comparison with SIN Task-specific method on the SIN'15 dataset, in terms of mean XinfAP.

	SIN 2015
Proposed MSCRs-related video retrieval approach (concept name + Semi-Aut. Augm.; no training exemplars)	0.2001
[9] (using annotated exemplars for training)	0.2630

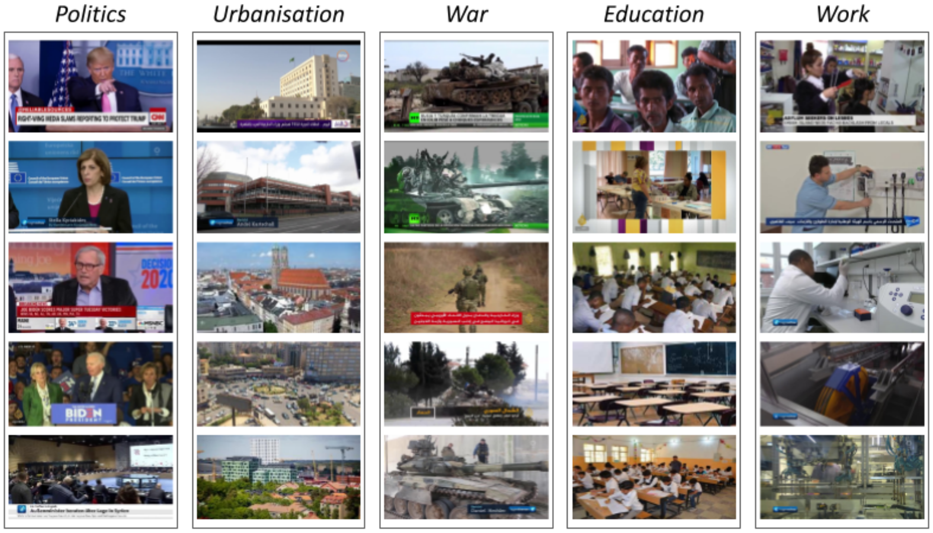

Figure 2: Example results, for five selected MRSCs. The top-5 retrieved images or video shots for each MRSC are shown.

4 EXPERIMENTS AND RESULTS

We evaluate the performance of our method using two datasets, i) the MIGRATION-VISUAL dataset as described in Section 3.3 and ii) the TRECVID 2015 Semantic Indexing task (SIN) dataset [11], for allowing comparison with learning-based video retrieval methods. The purpose of the TRECVID SIN task is to annotate video shots with concepts. In our experiments with the SIN'15 dataset, these concepts take over the position of the MRSCs [5]. The extended inferred average precision (XinfAP) [14] is used in both cases as the evaluation measure. XinfAP is an approximation of the mean average precision suitable for the partial ground-truth that accompanies the TRECVID dataset.

Our dual encoding deep network is trained using the combination of two large-scale video datasets: MSR-VTT [13] and TGIF [8]. Video shots' keyframes are initially represented utilizing a ResNet-152 deep network trained on the ImageNet-11k dataset. Also, two different word embeddings are used: i) Word2Vec [10] [4]; and, ii) BERT [3]. Finally, at the inference stage, the trained dual encoding deep network is used to create joint feature representations for the video shots/images of the evaluation datasets and for the MRSCs/SIN concepts as well as their augmentations.

In Table 3 the results on SIN'15 are presented. We can see that the automatic and semi-automatic augmentation approaches outperform the baseline that uses only the concept name as input. Also, their performance is very similar to that of [5], where human expert judges manually generated the augmentations. We noticed that following the automatic concepts augmentation i.e., the *"Concept name + Aut. Augm."* column, boosts the performance of the network. The augmentation procedure gives the necessary extra information the network needs to produce more accurate textual encoding. For that reason, it is easier for the network to associate concepts with the target video shots. In some cases, concepts that failed when only the concept name was used, e.g., *Demonstration Or Protest* or *Instrumental Musician*, exhibit greatly improved results. Moreover, human experts' involvement can further increase the performance for some concepts. However, since the experts are allowed only to delete some produced augmentations, in the SIN'15 dataset which is already mostly focused on visual concepts, the performance gain is limited.

In Table 4, the results on the MIGRATION-VISUAL dataset for ten selected MRSCs are presented. Our method's performance improved from 0.1554 to 0.4164 when the MRSCs were augmented automatically, compared to using only the MRSCs concept names. Further improvement is achieved in the semi-automated augmentation, with the mean XinfAP increasing to 0.5292. It is clear that when only the MRSC's name is used as input, some MRSCs perform reasonably well, e.g., *War*, *Politics* etc. while others like *Laws* or *Ethnic identity* fail. We noticed that the best-performing MRSCs are often closely-related to visible concepts (e.g. "Urbanization" relates to buildings and city skylines), which helps the network retrieve related videos or images. In contrast, when it comes to MRSCs with more abstract meanings, the network fails to retrieve related content. In all cases, the automatic augmentation significantly boosts the network's performance. Finally, human experts' intervention gives a further boost to very abstract and general MRSCs, i.e., *Laws* and *Life*, because they discard marginally-relevant augmentations and allow focusing on those that clearly define the MRSCs within the migration context. In general, semi-automatic augmentation is helpful in most cases.

To further illustrate our approach's performance, Figure 2 shows visual examples of the retrieved images or video shots for five example MRSCs: the top-5 results for each MRSC are presented. Finally, in Table 5 we compare our method on the SIN'15 dataset with a conventional concept retrieval method [9] that is trained with positive video exemplars for every concept. Even though our method does not outperform this supervised learning method, the goal of this comparison is to highlight that our approach performs relatively well, even though it does not need ground-truth concept-annotated images/videos for training.

5 CONCLUSION

We presented an image/video analysis and retrieval method that bridges the gap between abstract MRSCs and visual content without using any manually-generated MRSCs augmentations and without requiring a training corpus with (inevitably, manually-generated) ground-truth MRSC annotations. We showed, using not only a generic benchmark dataset but also a migration-related dataset that we introduced, that our method can effectively retrieve visual content relating to specific MRSCs under real-world conditions. The applicability of the proposed method is not limited to migration-related content; this line of work contributes to the general objective of developing efficient automatic methods for understanding how multimedia content that circulates in social media relates to complex semantic concepts, thus allowing us to study and better understand social phenomena.

ACKNOWLEDGMENTS

This work was supported by the EU Horizon 2020 programme under grant agreement 832921 (MIRROR).

REFERENCES

[1] George Awad, Asad A. Butt, Keith Curtis, Jonathan Fiscus, et al. 2020. TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVID 2020*. NIST, USA.

[2] Gerhard Backfried, Christian Schmidt, Mark Pfeiffer, Gerald Quirchmayr, and Johannes Göllner. 2015. Open Source Intelligence for Traditional- and Social Media Sources - The Sail Labs Media Mining System for OSINT. *The 10th International Conference iNCEB2015* (2015).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[4] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. 2018. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia (TMM)* 20, 12 (Dec 2018), 3377–3388.

[5] Erick Elejalde, Damianos Galanopoulos, Claudia Niederée, and Vasileios Mezaris. 2021. Migration-Related Semantic Concepts for the Retrieval of Relevant Video Content. In *Intelligent Technologies and Applications: Third International Conference, INTAP 2020*. Springer International Publishing, 404–416.

[6] Katie Eriksson. 2010. Concept determination as part of the development of knowledge in caring science. *Scandinavian Journal of Caring Sciences* 24 (2010), 2–11.

[7] Damianos Galanopoulos and Vasileios Mezaris. 2020. Attention Mechanisms, Signal Encodings and Fusion Strategies for Improved Ad-hoc Video Search with Dual Encoding Networks. In *Proceedings of the ACM International Conference on Multimedia Retrieval* (Dublin, Ireland) *((ICMR '20))*. ACM.

[8] Yuncheng Li, Yale Song, Liangliang Cao, et al. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4641–4650.

[9] Foteini Markatopoulou, Anastasia Ioannidou, Christos Tzelepis, et al. 2015. ITI-CERTH participation to TRECVID 2015. In *Proceedings of TRECVID 2015*. NIST, USA.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, Workshop Track Proceedings (ICLR '13)*.

[11] Paul Over, Jonathan Fiscus, David Joy, et al. 2015. TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2015*. NIST, USA.

[12] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference EMNLP*. Association for Computational Linguistics.

[13] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5288–5296.

[14] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st International ACM SIGIR 2008*. 603–610.