TOWARDS A BETTER UNDERSTANDING OF MODEL-FREE SEMANTIC CONCEPT DETECTION FOR ANNOTATION AND NEAR-DUPLICATE VIDEO CLIP DETECTION

Hyun-seok Min, Jae Young Choi, Wesley De Neve, and Yong Man Ro

Image and Video Systems Lab, Korea Advanced Institute of Science and Technology (KAIST),
Yuseong-gu, Daejeon, 305-701, Republic of Korea
{hsmin, jygchoi, wesley.deneve, ymro}@kaist.ac.kr

ABSTRACT

Given the observation that content transformations tend to preserve semantic information, we demonstrated in previous research that model-free semantic concept detection can be successfully leveraged for identifying NDVCs. In this paper, we seek a better understanding of the usefulness of model-free semantic concept detection for both the task of annotation and NDVC detection. In particular, through extensive experiments, we demonstrate that the problem of detecting semantic concepts for the goal of identifying NDVCs is more relaxed than the problem of detecting semantic concepts for annotation purposes: whereas incorrectly detected semantic concepts negatively affect the effectiveness of annotation, they do not negatively affect the effectiveness of NDVC detection, as long as the same incorrect semantic concepts are detected for both the reference and near-duplicate video clips. This observation has practical implications for the design of a video management system that makes use of model-free semantic concept detection for both the purpose of annotation and NDVC detection.

Index Terms— Annotation, folksonomy, near-duplicate video clip detection, semantic concept detection, video copy detection

1. INTRODUCTION

Near-duplicate video clip (NDVC) detection is at the core of multimedia applications such as media usage monitoring, content linking on the Web, metadata propagation for annotation purposes, protection of intellectual property, mitigation of visual redundancy in search results, and management of personal media libraries [1]. Given a reference video database and a query video clip, the goal of NDVC detection is to find all matches between the query video clip and the video clips in the reference video database. To that end, video clips are often represented by means of low-visual features extracted from keyframes. This representation is commonly referred to as a video signature. Low-level visual features may for instance describe color [2] or the spatial distribution of intensity information [3]. For an overview of content-based video copy detection, we would like to refer the reader to [4] and [5].

Given the observation that content transformations tend to preserve the semantic information conveyed by the altered video content [6], we demonstrated in previous research that semantic concept detection can be successfully leveraged for identifying NDVCs. In [7], we discuss the identification of NDVCs using model-based semantic concept detection, relying on the temporal variation of a limited number of semantic concepts to overcome a restricted concept vocabulary. In [8], we propose to identify NDVCs by using model-free semantic concept detection, exploiting an unrestricted concept vocabulary and eliminating the need for training. Specifically, we propose to take advantage of the collective knowledge in an image folksonomy (i.e., a collection of user-provided images and user-defined tags) for the goal of semantic concept detection and subsequent NDVC identification.

Whereas our previous research efforts focused on the effectiveness of NDVC detection, the aim of this paper is to seek a better understanding of the use of model-free semantic concept detection for both the task of annotation and NDVC detection. Based on experimental results obtained for the MIRFLICKR-25000 and the TRECVID 2009 datasets, we demonstrate that the problem of detecting semantic concepts for the goal of identifying NDVCs is more relaxed than the problem of detecting semantic concepts for annotation purposes. Specifically, we find that *incorrectly detected semantic concepts* do not negatively affect the effectiveness of NDVC detection, as long as the same incorrect semantic concepts are detected for both the reference and near-duplicate video clips. This observation has practical implications for the design of a video management system that makes use of model-free semantic concept detection for both the purpose of annotation and NDVC detection, and in particular for the selection of a feasible tag relevance threshold, used during the retrieval of images and tags from an image folksonomy.

This paper is organized as follows. Section 2 outlines our approach for model-free semantic concept detection, focusing on its use for identifying NDVCs. Experimental results are subsequently discussed in Section 3. Finally, conclusions and directions for future research are presented in Section 4.

2. NDVC DETECTION USING MODEL-FREE SEMANTIC CONCEPT DETECTION

Fig. 1 visualizes our technique for model-free semantic concept detection, and its subsequent use for annotation and NDVC identification purposes. In what follows, we focus on explaining our semantic approach for NDVC detection, given that this approach aims at taking model-free semantic concept detection a step further, compared to the task of annotation.

The following steps are used to determine whether a newly uploaded video clip is a near-duplicate version of a reference video clip: 1) video shot segmentation; 2) semantic concept detection; 3) creation of a semantic video signature; and 4) matching of semantic video signatures. The last three steps will be explained in more detail in the following subsections (see also [8]).

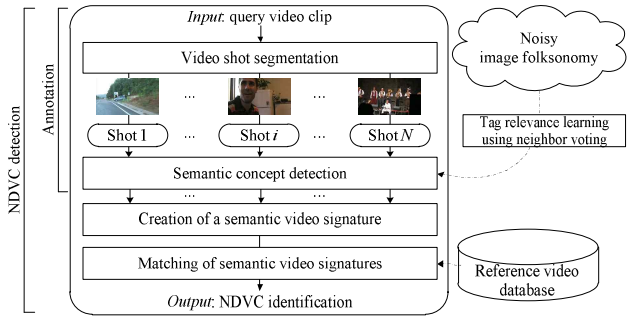

Fig. 1. Annotation and NDVC detection using model-free semantic concept detection.

2.1. Semantic concept detection

Semantic concept detection for a video clip $\mathbf{V}$ is performed at the level of shots. To that end, $\mathbf{V}$ is first segmented into N shots such that $\mathbf{V} = \langle \mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N \rangle$, with $\mathbf{S}_i$ denoting the i^{th} shot of $\mathbf{V}$. Next, each shot $\mathbf{S}_i$ is represented by a key frame, and a keyframe is in its turn represented by low-level visual features. The low-level visual features of the keyframe under consideration are subsquently used to retrieve the top K most visually similar images from an image folksonomy. The user-defined tags of the top K most visually similar images are retrieved as well. The relevance of a user-defined tag t to $\mathbf{S}_i$ is then measured as follows [9]:

$$R(t) = \frac{c}{K} - \frac{|L_t|}{|F|}, \tag{1}$$

where c denotes the frequency of t in the set of K visual neighbors of $\mathbf{S}_i$, where $|L_t|$ represents the number of images labeled with t in the folksonomy F, and where $|F|$ denotes the total number of images in F. If $R(t)$ is higher than a prespecified tag relevance threshold ζ_{tag}, then we assume that t is representative for $\mathbf{S}_i$.

2.2. Creation of a semantic video signature

Given that a shot usually contains multiple semantic concepts, we represent each shot by means of a *semantic feature signature* $\mathbf{A}_i$:

$$\mathbf{A}_i = \left[\left\langle t_{i,j}, w_{i,j} \right\rangle, j = 1,\ldots,\left|\mathbf{A}_i\right| \right], \tag{2}$$

where $t_{i,j}$ is the j^{th} semantic concept detected for $\mathbf{S}_i$. The weight value $w_{i,j}$ for $t_{i,j}$ is computed as follows:

$$w_{i,j} = \frac{R(t_{i,j})}{\sum_{k=1}^{|\mathbf{A}_i|} R(t_{i,k})}. \tag{3}$$

The semantic signature $\mathbf{U}$ of $\mathbf{V}$ is then defined as follows:

$$\mathbf{U} = \left\langle \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N \right\rangle. \tag{4}$$

2.3. Matching of semantic video signatures

Video matching aims at determining whether a query video clip appears in a reference video clip, and if so, at what location in the reference video clip. Given (4), we meassure the dissimilarity between a query video clip $\mathbf{V}^q$, having N shots, and a reference video clip $\mathbf{V}^r$, having L shots, as follows:

$$D(\mathbf{V}^q, \mathbf{V}^r) = \min_p \frac{1}{N} \sum_{i=1}^{N} D_{shot}(\mathbf{A}_i^q, \mathbf{A}_{i+p}^r), \tag{5}$$

where p denotes the position of the video shot in the reference video clip at which dissimilarity measurement starts. If $D(\mathbf{V}^q, \mathbf{V}^r)$ is smaller than a prespecified threshold ζ_{video}, then we assume that $\mathbf{V}^q$ is a near-duplicate of $\mathbf{V}^r$.

$D_{shot}(\mathbf{A}^q_i, \mathbf{A}^r_{i+p})$ in (5) represents the semantic dissimilarity between two video shots, measured using SQFD [10][11]:

$$\begin{aligned} D_{shot}(\mathbf{A}^q, \mathbf{A}^r) &= SQFD(\mathbf{A}^q, \mathbf{A}^r) \\ &= \sqrt{[\mathbf{W}^q \mid -\mathbf{W}^r]\mathbf{G}[\mathbf{W}^q \mid -\mathbf{W}^r]^T}, \end{aligned} \tag{6}$$

where $\mathbf{W}^q = [w_1^q, \ldots, w_{|\mathbf{A}^q|}^q]$ and $\mathbf{W}^r = [w_1^r, \ldots, w_{|\mathbf{A}^r|}^r]$ denote weight vectors, and where $[\mathbf{W}^q \mid -\mathbf{W}^r]$ denotes the concatenation of $\mathbf{W}^q$ and $-\mathbf{W}^r$. Furthermore, $\mathbf{G}$ denotes a ground similarity matrix of dimension $(|\mathbf{A}^q|+|\mathbf{A}^r|)\times(|\mathbf{A}^q|+|\mathbf{A}^r|)$. This matrix is the result of applying a similarity function to the semantic concepts assigned to the two video shots. The ground similarity between two semantic concepts is measured by means of tag (co-)occurrence statistics [12]. Specifically, the elements of $\mathbf{G}$ are computed as follows:

$$g_{ij} = \frac{\left|\mathbf{I}_{t_i \cap t_j}\right|}{\left|\mathbf{I}_{t_i}\right|}, \tag{7}$$

where the numerator denotes the set of images annotated with both t_i and t_j, and where the denominator denotes the set of images annotated with t. In our experiments, the tags t_i and t_j were used to query the Flickr image search engine in order to reliably estimate the number of images in the aforementioned sets.

As discussed in [10], SQFD bridges the gap between quadratic form distances and adaptive feature signatures, making it possible to take into account the fact that the nature and the number of detected semantic concepts may strongly vary from video shot to video shot. Moreover, SQFD is computationally efficient.

3. EXPERIMENTS

3.1. Experimental setup

Our experiments made use of the publicly available MIRFLICKR-25000 image set [13] as a source of collective knowledge. Also, our experiments relied on the publicly available TRECVID 2009 video set [14] to create a reference video database and NDVCs. TRECVID 2009 contains 400 video clips, having a total duration of 100 hours. Shot detection was performed by using the technique proposed in [15]. Further, the frame in the middle of each shot was used as a representative keyframe. In our experiments, Bag of Visual Words (BoVW) was used to represent the low-level visual content of folksonomy images and keyframes, using a vocabulary of 500 visual words, derived from 61,901 training images [16]. In addition, interest points were detected, described, and clustered using Difference of Gaussians (DoG), the Scale Invariant Feature Transform (SIFT), and k-means clustering, respectively.

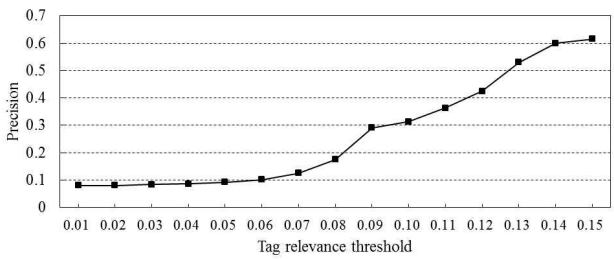

Fig. 2. Precision of annotation as a function of the tag relevance threshold.

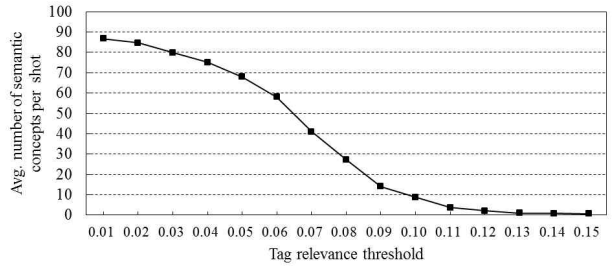

Fig. 3. Average number of detected semantic concepts per shot as a function of the tag relevance threshold.

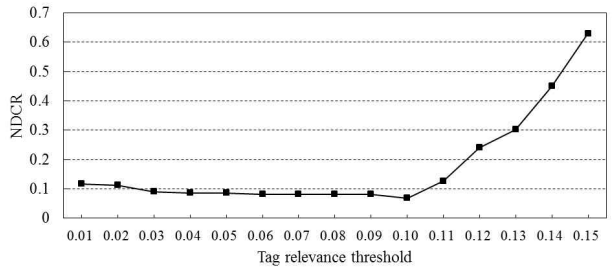

Fig. 4. NDCR as a function of the tag relevance threshold.

The Normalized Detection Cost Ratio (*NDCR*) was used to measure the effectiveness of NDVC detection [17]:

$$NDCR = P_{miss} + \beta \times R_{FA}, \tag{8}$$

where $$P_{miss} = \frac{N_{FN}}{N_{TP}},\ R_{FA} = \frac{N_{FP}}{T_{refdata} \times T_{query}}, \tag{9}$$

and P_{miss} is the probability of a miss and R_{FA} is the false alarm rate. Further, N_{TP}, N_{FN}, and N_{FP} denote the number of true positives, false negatives, and false positives, respectively, while $T_{refdata}$ and T_{query} represent the total duration of the reference video clips and the query video clips, respectively (duration is expressed in hours). In addition, β is a factor that trades off the cost of missing a true positive and the cost of having to deal with a false alarm. In our experiments, following the recommendation made by [17], we set $\beta = 2$.

Finally, the threshold for video matching ζ_{video} was empirically set to 0.4. Also, folksonomy-based semantic concept detection was realized using 10 nearest neighbor images (i.e., K is set to 10).

3.2. Experimental results

3.2.1. Effectiveness of annotation

We first investigate the influence of incorrectly detected semantic concepts on the effectiveness of annotation. To that end, we calculate the precision of the detected semantic concepts as a function of the tag relevance threshold ζ_{tag} (see Section 2.1), for 10 video clips randomly selected from the reference video database. Precision is computed by dividing the number of true positives per shot by the total number of true and false positives, and then averaging this number over all shots and video clips used. We assume that the precision of the detected semantic concepts for the 10 reference video clips is representative for their near-duplicate versions (given that transformations used to create NDVCs tend to preserve semantic information).

	Key frame	Detected semantic concepts (tag relevance threshold=0.01)	Detected semantic concepts (tag relevance threshold=0.10)	Detected semantic concepts (tag relevance threshold=0.15)
Reference		sky, night, star, geotagged, dark, nightscene, milky way, game, sea, constellation, impressedbeauty, concert, line, unitedkingdom, light, …	sky, night, star, geotagged, dark, nightscene, milky way, game, sea, constellation	sky, night
NDVC (Blur)		**star**, sagittarius, **milky way**, **sky**, **night**, moon, **sea**, eclipse, sunset, **light**, **impressedbeauty**, lunar, …	**star**, sagittarius, **milky way**, **sky**, **night**, moon, **sea**, eclipse, sunset	star, sagittarius, milky way
NDVC (Crop)		**night**, **milky way**, **sky**, moon, leamington, sagittarius, market, texture, galaxy, blue, **light**, …	**night**, **milky way**, **sky**, moon	**night**, milky way, **sky**
NDVC (Picture-in-picture)		**milky way**, **sky**, stars, **night**, aquila, scorpius, **constellation**, space, house, light, telescope, jupiter, …	**milky way**, **sky**, stars, **night**, aquila, scorpius	milky way, **sky**
NDVC (Brightness change)		**sky**, **night**, **star**, **geotagged**, **dark**, **nightscene**, **milky way**, **game**, **sea**, **constellation**, **impressedbeauty**, …	**sky**, **night**, **star**, **geotagged**, **dark**, **nightscene**, **milky way**, **game**, **sea**, **constellation**	**sky**, **night**

(a)

	Key frame	Detected semantic concepts (tag relevance threshold=0.01)	Detected semantic concepts (tag relevance threshold=0.10)	Detected semantic concepts (tag relevance threshold=0.15)
Reference		park, flower, animal, interestingness, cactus, grass, leaves, scenery, wild, animalplanet, africa, hunt, ….	park, flower, animal, interestingness	park, flower
NDVC (Blur)		**flower**, nature, **park**, spring, summer, **grass**, puppy, **wild**, tree, toy, …	**flower**, nature, **park**, spring	**flower**, nature
NDVC (Crop)		dog, **grass**, **park**, **flower**, **animal**, nature, **hunt**, **wild**, **scenery**, landscape, beautiful, art, **africa**…	dog, grass, **park**, flower, **animal**, nature	dog, grass
NDVC (Picture-in-picture)		nature, dog, **grass**, tree, **park**, **interestingness**, **africa**, toy, view, weddingday, cupcake, …	nature, dog, **grass**, tree, **park**, **interestingness**	nature
NDVC (Brightness change)		**park**, **flower**, **animal**, **interestingness**, **cactus**, **grass**, **leaves**, **scenery**, **wild**, **animalplanet**, **africa**, **hunt**, ….	**park**, **flower**, **animal**, **interestingness**	**park**, **flower**

(b)

Fig. 5. Example keyframes. Correct semantic concepts have been underlined. In addition, semantic concepts that have been detected for both the reference and near-duplicate video clips have been marked in bold.

Fig. 2 illustrates that the use of a higher tag relevance threshold results in an increase of the precision of annotation. This holds particularly true when the tag relevance threshold becomes higher than 0.10. However, even for a tag relevance threshold higher than 0.13, incorrectly detected semantic concepts are still present. Furthermore, Fig. 3 shows that the use of a higher tag relevance threshold results in a decrease of the average number of detected semantic concepts per shot (this average includes both true and false positives).

3.2.1. Effectiveness of NDVC detection

To investigate the influence of incorrectly detected semantic concepts on the effectiveness of NDVC detection, we created 40 NDVCs by applying four transformations (blurring, picture-in-picture, change in brightness, and cropping) to the 10 reference video clips used in Section 3.2.1.

Compared to the effectiveness of annotation, Fig. 4 shows that the effectiveness of NDVC detection is highly robust against a varying tag relevance threshold, as long as the threshold is between 0.01 and 0.11. When the threshold is higher than 0.11, the effectiveness of NDVC detection quickly decreases. This is due to the presence of a lower number of semantic concepts (see Fig. 3), resulting in less discriminative power.

Fig. 5 contains two example key frames, annotated with semantic concepts that have been detected by making use of the collective knowledge present in MIRFLICKR-25000. First, we can observe that several semantic concepts are irrelevant to the video content. This can mainly be attributed to two reasons: the presence of noisy tags in MIRFLICKR-25000 and the limitations of content-based retrieval [9][12]. Second, we can observe that the transformed keyframes have a significant number of detected semantic concepts in common with the unaltered keyframes, especially when the tag relevance threshold either has a value of 0.01 or 0.10. All of these common concepts contribute to a higher NDVC detection effectiveness, regardless of the fact whether these semantic concepts are correct or not.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we discussed the usefullnes of model-free semantic concept detection for both the task of annotation and NDVC detection. Our experimental results demonstrate that the problem of detecting semantic concepts for the goal of identifying NDVCs is more relaxed than the problem of detecting semantic concepts for annotation purposes: whereas incorrectly detected semantic concepts negatively affect the effectiveness of automatic annotation, they do not negatively affect the effectiveness of NDVC detection, as long as the same incorrect semantic concepts are detected for both the original and near-duplicate video clips.

The aforementioned observation is of practical significance for the design of a video management system that simultaneously aims at annotating newly uploaded video clips and detecting whether these video clips are near-duplicate versions of video clips in a reference video database. For a trade-off exists that influences the selection of a feasible tag relevance threshold, used during the retrieval of images and tags from an image folksonomy: the use of a high tag relevance threshold may result in a high precision of annotation, but in an NDVC detection effectiveness that is low (due to a high number of false positives), and vice versa.

In future research, we plan to investigate whether the presented NDVC detection system fullfills the requiremens of *stochastic resonance*, wherein adding noise to a threshold measurement system can sometimes improve its performance.

5. ACKNOWLEDGEMENTS

This research was supported by the Basic Science Research Program of the National Research Foundation (NRF) of Korea, funded by the Ministry of Education, Science and Technology (research grant: 2011-0011383).

6. REFERENCES

[1] P. Brasnett, S. Paschalakis, and M. Bober, "Recent developments on standardisation of MPEG-7 Visual Signature Tools," *Proc. of ICME*, pp. 1347-1352, 2010.

[2] S.H. Kim and R.-H. Park, "An efficient algorithm for video sequence matching using the modified Hausdorff distance and the directed divergence," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 7, pp. 592-296, July 2002.

[3] D. N. Bhat and S. K. Nayar, "Ordinal measures for image correspondence," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 415-423, 1998.

[4] J. Law-To, L. Chen, A. Joly, I. Laptev, and O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, "Video Copy Detection: a Comparative Study," *Proc. of CIVR*, Amsterdam, The Netherlands, July 9-11, 2007.

[5] A. Sarkar, V. Singh, P. Ghosh, B. S. Manjunath, and A. Singh, "Efficient and Robust Detection of Duplicate Videos in a Large Database," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 20, no. 6, pp. 870-885, 2010.

[6] R. D. Oliveira, M. Cherubini, and N. Oliver, "Looking at near-duplicate videos from a human-centric perspective," *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 6, no. 3, August 2010.

[7] H-S. Min, W. De Neve, and Y.M. Ro, "Towards using semantic features for near duplicate video detection", *Proc. of ICME*, pp. 1364-1369, 2010.

[8] H-S. Min, W. De Neve, and Y.M. Ro, "Exploiting Collective Knowledge in an Image Folksonomy for Semantic-Based Near-Duplicate Video Detection," *Proc. of ICIP*, pp. 3165-3168, 2010.

[9] X. Li, C. G. M. Snoek, and M. Worring, "Learning Social Tag Relevance by Neighbor Voting," *IEEE Trans. on Multimedia*, vol. 11, no. 7, pp. 1310-1322, 2009.

[10] C. Beecks, M. S. Uysal, and T. Seidl, "Signature Quadratic Form Distance," *Proc. of CIVR*, pp. 438-445, 2010.

[11] C. Beecks, M. S. Uysal, and T. Seidl, "Signature Quadratic Form Distances for Content-Based Similarity," *ACM Multimedia*, pp. 697-700, 2009.

[12] S. Lee, W. De Neve, and Y. M. Ro, "Tag Refinement in an Image Folksonomy using Visual Similarity and Tag Co-occurrence Statistics," *Signal Processing - Image Communication*, vol. 25, no. 10, 2010.

[13] M. J. Huiskes and M. S. Lew, "The MIR Flickr Retrieval Evaluation," *ACM International Conference on Multimedia Information Retrieval*, pp. 39-43, 2008.

[14] P. Over et al., "TRECVID 2009 - goals, tasks, data, evaluation mechanisms and metrics," In TRECVid Workshop 2009, November 2009.

[15] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System," TREC Video Retrieval Evaluation Online Proceedings, 2004.

[16] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," *Proc. of CIVR*, pp. 494-501, 2007.

[17] CBCD Evaluation Plan TRECVID 2010, available on http://www-nlpir.nist.gov/projects/tv2010/Evaluation-cbcd-v1.3.htm#eval.