An Automated End-To-End Pipeline for Fine-Grained Video Annotation Using Deep Neural Networks

Baptist Vandersmissen, Lucas Sterckx, Thomas Demeester, Azarakhsh Jalalvand, Wesley De Neve, and Rik Van de Walle

In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 409–412, 2016.

http://doi.acm.org/10.1145/2911996.2912028

An Automated End-To-End Pipeline for Fine-Grained Video Annotation using Deep Neural Networks

Baptist Vandersmissen[1] Lucas Sterckx[3] Thomas Demeester[3]

Azarakhsh Jalalvand[1] Wesley De Neve[1,2] Rik Van de Walle[1]

{firstname.lastname}@ugent.be

[1] Data Science Lab, ELIS, Ghent University – iMinds, Ghent, Belgium
[2] Image and Video Systems Lab, KAIST, Daejeon, South Korea
[3] Internet Based Communication Networks and Services, INTEC, Ghent University – iMinds, Ghent, Belgium

ABSTRACT

The searchability of video content is often limited to the descriptions authors and/or annotators care to provide. The level of description can range from absolutely nothing to fine-grained annotations at the level of frames. Based on these annotations, certain parts of the video content are more searchable than others.

Within the context of the STEAMER project, we developed an innovative end-to-end system that attempts to tackle the problem of unsupervised retrieval of news video content, leveraging multiple information streams and deep neural networks. In particular, we extracted keyphrases and named entities from transcripts, subsequently refining these keyphrases and named entities based on their visual appearance in the news video content. Moreover, to allow for fine-grained frame-level annotations, we temporally located high-confidence keyphrases in the news video content. To that end, we had to tackle challenges such as the automatic construction of training sets and the automatic assessment of keyphrase imageability.

In this paper, we discuss the main components of our end-to-end system, capable of transforming textual and visual information into fine-grained video annotations.

Keywords

deep neural networks, fine-grained video annotation, video retrieval

1. INTRODUCTION

Fully automated metadata generation for multimedia content is the holy grail of information retrieval, given that fine-grained annotation enables more accurate searchability in multimedia archives. In this paper, we present a fully automated end-to-end system for generating metadata for news video content. Indeed, television broadcasters have a great need for accurate annotations, so to ensure future searchability and monetization of their video content.

The system discussed in this paper was developed within the context of the STEAMER project[1], a collaborative research project funded by both industry and the Flemish government. Our system extracts relevant keyphrases and candidate persons from information residing in transcripts. To cope with the different nature of keyphrases (concept detection) and candidate persons (face recognition), our systems processes the two types of information differently. The relevance of a subset of keyphrases is reassessed based on the presence of visual elements linked to the core concepts of each specific keyphrase. Candidate persons (discovered via named entity extraction) are visually matched based on deep face features to increase the relevance and ranking of these annotations. Both types of information are processed by making use of state-of-the-art deep learning techniques, for example to extract effective features.

Furthermore, the system discussed in this paper supports the learning and recognition of an infinite number of concepts, thus making it possible to move away from a limited semantic coverage. To do so, we developed methods to automatically construct training sets and to automatically assess the visual representativeness of keyphrases. In addition, our system allows for fine-grained annotation of news video content with both general and specific concepts (via keyphrases), as well as with the presence of noteworthy persons (via named entity detection), thus greatly increasing the searchability of the news video content in question.

After a brief review of related work in Section 2, we describe the main components of our system in Section 3. Next, we discuss both quantitave and qualitative results in Section 4. Finally, we conclude our work in Section 5, also giving some hints for future improvements.

2. RELATED WORK

The problem of automatic multimedia enrichment has been tackled by many projects before. In this section, we dis-

[1] http://www.iminds.be/en/projects/2014/07/12/steamer

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912028

cuss two recent projects that are of interest, namely AXES [10] and LinkedTV [8]. The AXES project[2] attempted to develop tools that provide various types of users with new ways to interact with audiovisual libraries. Specifically, information sources such as audio streams and blogs were used to increase the overall explorability of audiovisual content. LinkedTV[3] is a similar project that uses semi-automatic audiovisual analysis of television content to generate links to relevant web content.

The essential difference between these projects and the system proposed in this paper is the fact that our system attempts to enrich the video in an unconstrained manner without any human interference. This means that our sytem considers every possible relevant keyphrase, introducing challenges such as the automatic building and cleaning of training sets, as well as the assessment of the visual representativeness of a keyphrase. Regarding the automatic building of training sets, the authors of [2, 3, 5, 6] introduced novel ideas to collect and increase the overall relevancy of both the positive and negative training samples, whereas the authors of [9] provided an in-depth analysis of the concept of visual representativeness.

3. METHODOLOGY

In this section, we describe the overall architecture of the proposed system and we explain the different components built. Figure 1 provides a visual overview of these components. As depicted, a video segment and an accompanying transcript are used as inputs. Next, after processing, the extracted transcript information is matched against the visual content, making it possible to adjust the relevance of keyphrases and candidate persons.

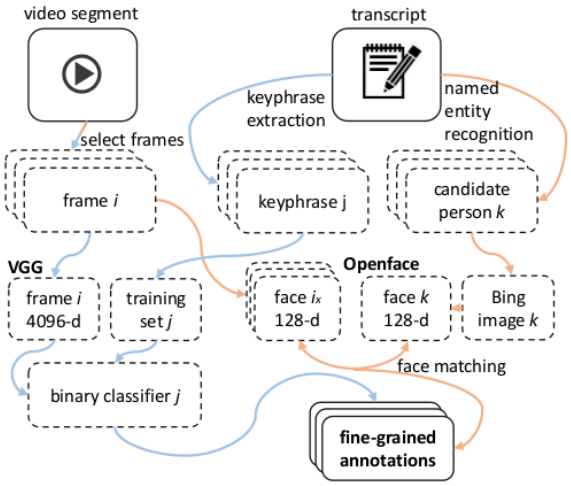


Figure 1: Simplified overview of our system.

In what follows, we describe the general flow of our system in more detail, outlining the role and functioning of the different components.

3.1 Transcript Information

Transcript information is reduced to extracted keyphrases and candidate persons.

[2]http://www.axes-project.eu/
[3]http://www.linkedtv.eu/

3.1.1 Keyphrase Extraction

Candidate keyphrases are extracted by Part-Of-Speech (PoS) tagging the transcript and by filtering sequences of words with tags satisfying a regular expression that takes the form of <JJ>*<NN.*>+. The PoS tag JJ refers to an adjective, while NN refers to a noun. Keyphrases are then ranked according to a supervised model trained on annotated documents provided by the VRT, a Flemish broadcaster and one of the project partners. In particular, 1,250 documents were annotated with keyphrases by 300 users instructed to select the most prominent phrases in the document. Each document was assigned six keyphrases on average, and where these keyphrases were used as positively labeled training samples; candidates not in this set were assigned a negative label. For each candidate, the following features were extracted: (i) keyphrase frequency; (ii) number of tokens in the keyphrase; (iii) length of the longest term in the keyphrase; (iv) a binary feature indicating whether the keyphrase contains a named entity; (v) PoS tags; (vi) relative position of the first and last occurrence; (vii) span (relative last occurrence minus relative first occurrence); and (viii) TF*IDF score. The supervised model was trained by means of a binary classifier using logistic regression.

3.1.2 Collection of Images

Since we aim for a fully automated system with support for an unlimited number of keyphrases, several challenging steps are introduced. This includes filtering of keyphrases based on their visual relevance (imageability) and the automatic construction of training sets. The top j ranked keyphrases are taken into account as possible accurate descriptors for the linked video segment under consideration. Each of these keyphrase is then automatically translated from Dutch to English using Google Translate. Next, a set of representative images is collected using several commercial and social image search engines (that is, Bing Images, Google Images and Flickr). Challenges like wrongly translated, ambiguous or rare keyphrases increase the degree of difficulty to obtain representative images.

3.1.3 Imageability of a Keyphrase

A key difficulty is the assessment of the visual nature of a keyphrase (i.e., the imageability of a keyphrase). Tag imageability or representativeness can be defined as the extent to which a tag is able to effectively describe the visual content of a set of images annotated by that tag [9]. In this context, a tag is imageable if images annotated with that tag share similar concepts (that is, objects, scenes and actions).[4]

Assessing the imageability of a keyphrase is an important step as numerous keyphrases do not depict a single concept that is clearly visual in nature. Examples are abstract concepts like *"thoughts"* and *"interests"*, *feelings* and *notions of time*. These examples are nonvisual in nature, thus making it difficult or merely impossible to recognize these examples in video content. Therefore, we developed an algorithm to evaluate the imageabilty of a keyphrase.

This algorithm analyses a number of properties of the collected set of images: (i) the average Euclidean distance between all images based on a feature representation (cf. Section 3.1.4); (ii) the number of outliers in the image set; and

[4]The concept of imageability comes with a dubious interpretation. However, we consider a more detailed treatment of this topic to be out-of-scope for this demo paper.

(iii) the number of clusters based on a hierarchical clustering algorithm. A Support Vector Machine (SVM) model is trained based on a set of 200 keyphrases annotated with a binary label.

3.1.4 Training of a Visual Model

To accurately recognize a keyphrase in a video segment, a visual model for each keyphrase is trained. To train such a model, a training set is automatically collected. Positive samples are collected via aforementioned search engines and filtered for outliers. An outlier image is identified based on its overall distance to all other images in the set. Negative samples are collected by randomly selecting the same number of samples from a fixed collection of images (depicting typical news content) and images originating from previously collected and differing keyphrases.

Next, for each image, an effective feature representation is computed based on recently developed deep convolutional neural networks. In that regard, we use the popular framework Caffe [4] to extract the activations of layer FC_7 from the state-of-the-art Visual Geometry Group (VGG) 16-layer network [7] trained on the ImageNet Challenge dataset[5]. Each image is preprocessed (i.e., resized, cropped and mean subtracted) and feed forwarded through the network, resulting in a 4096-dimensional feature vector.

Finally, a model based on a linear SVM is trained using the extracted feature vectors as input, combined with a dedicated target indicating whether the image is either positive or negative with respect to the target keyphrase. This model is used as an indicator for the visual presence of a keyphrase in a video segment at a specific timestamp.

3.1.5 Candidate Persons

Named entities, and candidate persons in particular, are extracted from transcripts using software developed by the company Zeticon[6], also a STEAMER project partner. Five representative images are downloaded for each candidate person using the top results of the above described search engines (by filtering on portrait images only). The face in each image is detected and aligned. Next, a 128-dimensional representation is obtained using the activations of the last layer of the Openface deep neural network [1]. This representation contains different core characteristics of a face, enabling a cross-age comparison between multiple representations. Two faces are matched based on the L_2 distance between their face representations. This implies a candidate person is deemed to be visually present in the video if its representative face has been matched to one or more detected faces in the video content.

3.2 Video Segments

The video segments used as an input to our system are part of daily television news broadcast episodes. They contain in-studio news announcements, interviews, debates and reports.To be able to process video segments in a reasonable amount of time, the video is downsampled to one frame every second, thus discarding temporal information (cf. Section 5 for future work). Following the approach previously described in Section 3.1.4, for each of these frames, the activations of the FC_7 layer of the VGG deep convolutional neural network are extracted and used as a representation.

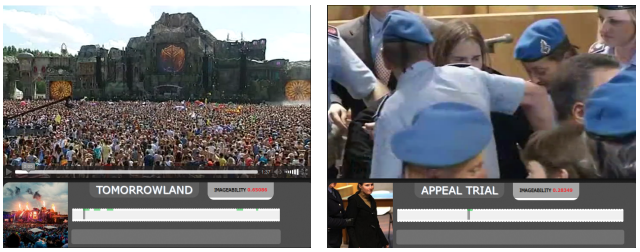
[5]http://www.image-net.org/challenges/LSVRC/
[6]http://www.zeticon.com/

3.3 Workflow

To automatically enrich a video segment with fine-grained annotations, the representation of each frame is fed as an input to all keyphrase classifiers separately. The output of each keyphrase classifier is a measure as to what extent the keyphrase is visually present in the video content. Further, all outcomes are aggregated over all frames to generate a final score per keyphrase, denoting its visual presence.

Regarding the recognition of candidate persons in a video segment, all face patches are first detected per selected frame. This is done by applying a combination of the Viola-Jones algorithm [11] (detection) and a deep convolutional neural network (confirmation). For each detected and aligned face patch, the corresponding face representation is then compared to all representations retrieved for the candidate persons. Depending on a threshold, video segment face patches are linked to candidate persons.

The final output of our system consists of a video segment with both coarse- and fine-grained textual keyphrase annotations, filtered and reranked based on their visual presence in the video. In addition, the video segment in question has been annotated with visually occurring persons on a frame-level basis.



(a) (b)

Figure 2: Example metadata generation for keyphrases "Tomorrowland" and "Appeal Trial".

4. RESULTS

To evaluate the overall performance of the proposed system, we need frame-level annotations for all considered keyphrases and candidate persons. Due to the current absence of such a dataset, we mainly provide a qualitative evaluation of the different functionalities of the proposed system.

4.1 Keyphrases

For each video segment, the top-k best scoring keyphrases are taken into consideration as possible video enrichments.

4.1.1 Imageability

We assess the correctness of the imageability scores attached to a number of keyphrases based on a constructed test set. Specifically, we randomly selected 100 keyphrases originating from a large set of transcript files, having these annotated by at least five different annotators. Given a clear definition of imageability, each keyphrase is assigned a label $\in$ {"*Yes*", "*Mostly Yes*", "*Mostly No*", "*No*"}. These labels are subsequently transformed into a weight $\in$ {1.0, 0.67, 0.33, 0} and averaged over all annotators resulting in one imageability score per keyphrase. The annotator agreement is measured by averaging the F_1 scores over all annotators (leaving the remaining annotators as ground truth), resulting in a score of 69%. Our predictor achieves a precision

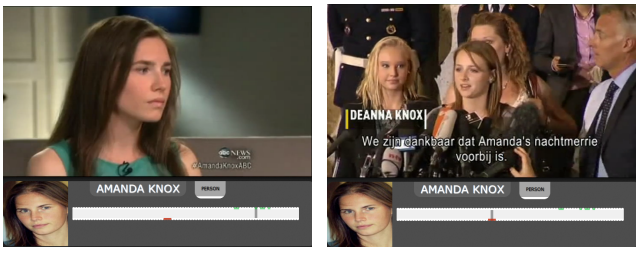
of 47% and a recall of 64% on this test set, resulting in an F_1 score of 55%. Considering the challenging nature of this task, we state that the difference between both F_1 scores (annotator versus algorithm) is reasonably small, showing the effectiveness of our approach.

4.1.2 Fine-grained Enrichments

The examplary Figure 2(a) shows the correct identification of the keyphrase "Tomorrowland". However, going from transcript data to relevant and imageable keyphrases introduces many hazards. Most mistakes are due to keyphrases depicting concepts that are either too general or too specific, resulting in a training set primarily consisting of noisy and diverse images. This is something that can be observed in Figure 2(b). In this example, the keyphrase "appeal trial" is to be recognized. Whereas the concept of a "trial" itself depicts a clear visual concept, the addition of "appeal" greatly increases the abstract nature of this keyphrase.

4.2 Candidate Persons

The recognition of faces and thus the presence of certain people in video content is challenging. This is primarily due to the large variation of representative images, the different styles, and the different age people can have in video segments. The main difficulties are a lack of online representative images for unknown persons and candidate persons being matched to lookalikes in the video. An example of such conflict is shown in Figure 3. The system correctly identifies "Amanda Knox" in Figure 3(a) but generates a false positive by assuming "Deanna Knox" to be her sister in Figure 3(b).



(a) Correct. (b) Incorrect.

Figure 3: Metadata generated for "Amanda Knoxx".

5. CONCLUSIONS & FUTURE RESEARCH

In this paper, we have introduced a fully automated end-to-end pipeline for generating fine-grained metadata for news broadcast video segments. Our discussion mainly focussed on the automatic construction of training sets and the automatic assessment of the visual representativeness of a keyphrase. In that regard, we have trained a linear SVM-based classifier for each keyphrase, leveraging features extracted from a deep convolutional neural network. Some qualitative evaluation of our system indicates that the proposed system could greatly aid broadcasters in organizing their content, increasing its overall searchability and monetization.

In future research, we will focus on an extensive quantitative analysis of the effectiveness of our system. In addition, we plan to investigate methods to improve the automatic construction of training sets, given that empirical evidence has shown this to be a crucial step in our pipeline. Furthermore, we will improve the accuracy of our algorithm for tag imageability assessment. Finally, we will study several aggregation methods in more detail, as well as the inclusion of temporal information, so to facilitate more effective keyphrase models.

6. ACKNOWLEDGEMENTS

The research presented in this article relates to STEAMER, a MiX-ICON project facilitated by iMinds Media and funded by IWT and Innoviris.

7. REFERENCES

[1] B. Amos, B. Ludwiczuk, J. Harkes, P. Pillai, K. Elgazzar, and M. Satyanarayanan. OpenFace: Face Recognition with Deep Neural Networks. http://github.com/cmusatyalab/openface.

[2] K. Chatfield, R. Arandjelović, O. M. Parkhi, and A. Zisserman. On-the-fly learning for visual search of large-scale image and video datasets. *International Journal of Multimedia Information Retrieval*, 2015.

[3] N. Ikizler-Cinbis, R. Cinbis, and S. Sclaroff. Learning actions from the web. In *12th International Conference on Computer Vision IEEE*, pages 995–1002, 2009.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[5] S. Kordumova, X. Li, and C. G. M. Snoek. Best practices for learning video concept detectors from social media examples. *Multimedia Tools and Applications*, pages 1291–1315, 2014.

[6] X. Li, C. Snoek, M. Worring, D. Koelma, and A. Smeulders. Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia*, pages 933–945, 2013.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[8] D. Stein, E. Apostolidis, V. Mezaris, N. de Abreu Pereira, and J. Müller. Semi-automatic video analysis for linking television to the web. In *FutureTV Workshop*, pages 1–8, 2012.

[9] A. Sun and S. Bhowmick. Quantifying visual-representativeness of social image tags using image tag clarity. In *Social Media Modeling and Computing*, pages 3–23. 2011.

[10] P. Van Der Kreeft, K. Macquarrie, M. Kemman, M. Kleppe, and K. McGuinness. Axes-research 2014; a user-oriented tool for enhanced multimodal search and retrieval in audiovisual libraries. In *12th International Workshop on Content-Based Multimedia Indexing*, pages 1–4, 2014.

[11] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, pages 511–518, 2001.