Multimedia Analysis of Video Sources

Naiara Aginako[1], Juan Arraiza Irujo[1], Montse Cuadros[1], Matteo Raffaelli[2], Olga Kaehm[3], Naser Damer[3], Joao P. Neto[4]

[1] *Vicomtech-IK4, Paseo Mikeletegi 57, 20009 Donostia-San Sebastin, Spain*

[2] *Synthema, 56121 Pisa, Italy*

[3] *Fraunhofer Institute for Computer Graphics Research (IGD), 64283 Darmstadt, Germany*

[4] *VoiceInteraction, 1000-029 Lisbon, Portugal*

jarraiza,mcuadros,naginako@vicomtech.org, matteo.raffaelli@synthema.it, olga.kaehm,naser.damer@igd.fraunhofer.de, joao.neto@voiceinteraction.pt

Keywords: Multimedia analysis, Video Processing, Multimodal Surveillance, Decision Support Systems

Abstract: Law Enforcement Agencies (LEAs) spend increasing efforts and resources on monitoring open sources, searching for suspicious behaviours and crime clues. The task of efficiently and effectively monitoring open sources is strongly linked to the capability of automatically retrieving and analyzing multimedia data. This paper presents a multimodal analytics system, created in cooperation with European LEAs. In particular it is described how the video analytics subsystem produces a workflow of multimedia data analysis processes. After a first analysis of video files, images are extracted in order to perform image comparison, classification and face recognition. In addition, audio content is extracted to perform speaker recognition and multilingual analysis of text transcripts. The integration of multimedia analysis results allows LEAs to extract pertinent knowledge from the gathered information.

1 INTRODUCTION

The digital age has produced unparalleled access to a proliferation of multimedia data. A huge amount of multimedia data (video, image, audio and text contents) is now available in open sources. The digital age has also facilitated the growth of organized crime in the same way that it has reduced barriers for enterprises. The richness and quantity of information available from open sources, if properly gathered and processed, can provide valuable intelligence. That is the reason why Law Enforcement Agencies (LEAs) are becoming more inclined to using open source intelligence (OSINT) tools.

This paper presents an OSINT system, created in cooperation with European LEAs, for the detection and prevention of organized crime. The system is able to gather and process raw open source data, heterogeneous both for source, format, protocol and language, with the aim of retrieving social networks. These are finally explored using Visual Analytics (VA) technologies. In particular, we focus here on the video analytics subsystem. We describe the workflow of multimedia data analysis processes that is generated by the system once a video has been crawled.

The system crawler allows to retrieve multimedia contents in four ways: (1) Looking for a parametric number of documents on the web with a key-words search; (2) Looking for documents in a given Uniform Resource Locator (URL) until a parametric depth of levels, based on specified key-words; (3) Crawling Online Social Networks (OSNs), in particular Facebook's User Generated Contents (UGCs), focusing on the information contained in posts (textual content, photos, videos, creation time, etc.); (4) capturing all of the free-to-air signals (television and radio programs) that are received in a specific location or environment. Once a video has been crawled (either from the web, from Facebook or from a tv program), it is reduced to its base components and the following analysis workflows are automatically launched:

1. Image analysis: Images are extracted and compared with a set of reference images, providing a similarity score and a class belonging probability.
2. Face recognition: Automatic face recognition services are applied to the images of a face available in a video stream for person identification and verification.
3. Audio analysis: Audio content is extracted and

processed for speaker recognition and tracking, gender and age identification, etc.

4. Text analysis: Multilingual text analysis is applied to the text transcripts of the extracted audio content. Natural Language Processing (NLP) techniques are used to retrieve entity relationships.

In this paper, we will begin by giving a quick overview of currently available solutions for multimedia analysis. We will then outline the system architecture, focusing on the different analysis workflows mentioned above.

2 RELATED WORK

Current solutions regarding the analysis of video, image, audio and text contents from open sources are not reported here in detail. Regarding complete video analysis there are known projects such as Kinesense-VCA[1], where advanced video analysis technologies are implemented in order to help investigators in detecting suspicious behaviors in videos. Another relevant solution for video surveillance in the security domain is the one offered by IKUSI[2]. In addition, current European projects such as SAVASA[3] or MOSAIC[4] are developing solutions for the creation of video archive search and analysis platforms that also include semantic tools.

The variety of image analysis platforms is huge. The most famous and widely used one among them is Google Image Search, that introduced in 2011 the Search by Image functionality[5]. Other widely used applications are TinEye[6] and CamFind[7], that allow to find different kinds of information about the input image.

Regarding text analysis, many platforms are available. Most of these platforms perform textual analysis in only a few languages, and only a few of these platforms are endowed with crawling or visualization capabilities. Additionally, they usually do not include specific knowledge integrated such as domain ontologies. The most common platforms are AlchemyAPI[8], GATE[9], Lexalytics[10] or LanguageWare[11]. An interesting approach is Attensity[12], which relies mostly on social media sources. The most similar text analysis tool to our proposal is the one from BasisTech[13] covering up to 55 languages, using Google technology.

All these platforms tackle the analysis of diverse multimedia contents as independent solutions, but the real challenge appears to be the creation of a unique platform that integrates the most important multimedia analysis technologies. As aforementioned, the presented system encompasses all the analysis methods necessary for an integrate processing of the input multimedia streams. The capability of extracting interrelated knowledge from the input data allows LEAs to significantly reduce data collection, evaluation and integration efforts.

3 METHODOLOGY

3.1 Architecture description

The OSINT system described in this paper includes several distributed components. The back-bone of the system is an Enterpise Service Bus (ESB), which has been configured to allow integration, communication, and orchestration of all those independent but mutually interacting components. These components are loosely coupled and they integrate using Service Oriented Architecture (SOA) services. Each component has published a description of the web service and operations it offers using the Web Services Description Language (WSDL).

It is worth mentioning that the components might run in different operating systems, might have been developed in different programming languages, and might use internally as many repositories or other resources as needed. The overall system relies on each component to implement and offer what has been defined in its WSDL. As long as those services and operations respond well, the rest of the system does not need to know how that work has been implemented. One benefit of this approach is that it allows replacing one component by another one as long as the new one complies with the defined integration interface, in other words, as long as it complies with the defined WSDL.

[1] `http://www.kinesense-vca.com/product/`

[2] `http://www.ikusi.com/en/sport/solutions/security/video-surveillance-and-advanced-video`

[3] `http://www.savasa.eu/`

[4] `http://www.mosaic-fp7.eu/`

[5] `http://www.google.com/insidesearch/features/images/searchbyimage.html`

[6] `https://www.tineye.com/`

[7] `http://camfindapp.com`

[8] `http://www.alchemyapi.com/`

[9] `http://gate.ac.uk/`

[10] `http://www.lexalytics.com/`

[11] `http://www-01.ibm.com/software/ebusiness/jstart/downloads/LanguageWareOverview.pdf`

[12] `http://www.attensity.com/home/`

[13] `http://www.basistech.com/`

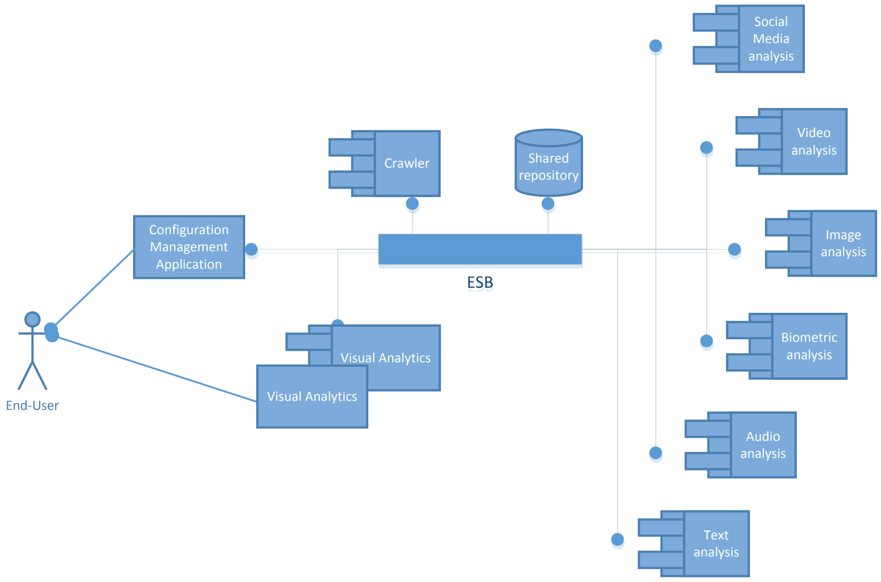

Figure 1: High level view of the architecture

Another important component is the common (or shared) repository. This repository is a Not Only Structured Query Language (NoSQL) repository and has three logical views. In the first one the original version of the crawled content is stored; this view is known as the original repository. A different logical view stores the normalized version of the crawled content. The crawler converts each content type to one of the accepted normalized formats that the analysis components can process; this second view is known as the normalized repository. Finally, a third view stores the output of the analysis modules; this view is known as the knowledge repository. Besides the analysis components, the visual analytics component also accesses this repository third view.

3.2 Video analysis

As aforementioned, the video analytics subsystem is the starting point of the analysis process. In the last years, the retrieval of highly representative information from videos has become one of the main topics of the research community. Moreover, automatic content analysis is very important for the efficient interpretation of all the information encompassed in the multimedia stream (Maybury, 2012). The consideration within the project of video files as the highest representative of the compilation of diverse contents engenders the necessity for the development of a video analysis module.

The first step for the consumption of the video content is the demultiplexation of the diverse nature tracks and their transcodification to more appropriate formats in order to ease the analysis. These considered tracks are classified as text, audio and images which represent the input units for the other analysis modules deployed in the project. While audio and text contents are directly obtained after the demultiplexation of the video file, the extraction of the relevant images for the analysis constitute a trade-off. The implemented Shot Boundary Detection (SBD) (Zhang et al., 2012)(Zhang et al., 1993) and Best Frame Extraction (BFE) (Rui et al., 1998)(Ejaz et al., 2012)(Sun and Fu, 2003) approaches tackle this problem.

Video content segmentation or structuring has been defined as the hierarchical decomposition of videos into units. In this solution, shots are considered as the representative units of the video content, so SBD detectors have been implemented to extract these basic units. Two different detectors are available in the platform, the first based on contiguous frames histogram similarity analysis and the second based on Discrete Cosine Transform (DCT) components calculated for different compression and codification purposes. Once the shots are identified, one frame for each shot is selected for the further image analysis to dismiss the high data redundancy in videos, which slows down the analysis process. The frame is selected considering the information contained within each shot. To that end, a feature vector is calculated for all the frames and the one with the minimum dis-

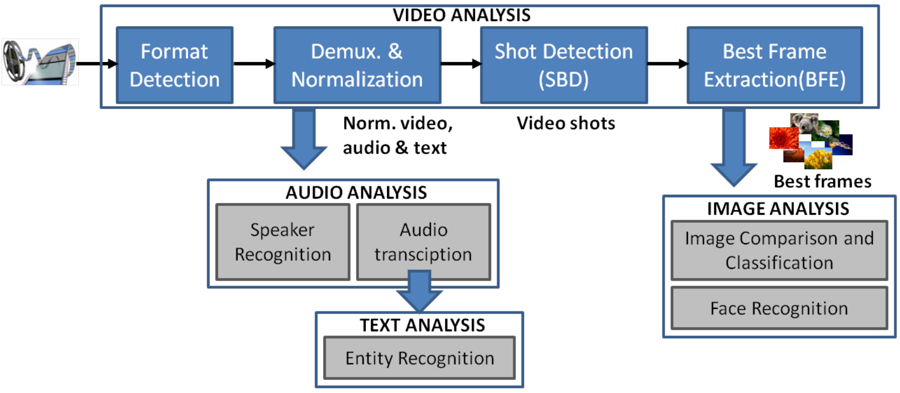

Figure 2: High level view of the video analytics subsystem architecture

tance to the rest of the frames is extracted.

Once the information of the video content has been separated, more specific analysis for each type of content is invoked to give the platform the most concise information about the multimedia content.

3.3 Image analysis on video frames

Once a video has been processed, the most significant videoimages (best frames) are extracted and saved in the normalized repository of the system, becoming available for the image analysis module.

There are two functionalities related with image analysis, image comparison and image classification. As mentioned above in 3.1, before launching a Research Line (RL), the user configures the system and can introduce among other parameters different sets of images. One kind of set represents images that the user wants to detect in collected data by image comparison, and the other one is used as the training set for the classification process. All the images belonging to the same classification set contain objects of the same class.

Once video analysis has been performed and the best frames have been uploaded into the normalized repository, the orchestrator calls the image analysis module to process them. First of all, feature points and their descriptors are extracted for each image (Tuytelaars and Mikolajczyk, 2008)(Bay et al., 2006)(Lowe, 2003). Depending on the process function, the module loads the RefModel from the Knowledge Repository which contains the descriptors or the class model of the already preprocessed Reference Sets. Afterwards, a comparison between the input image and the RefSet images or the classification of the input image is performed (Chen et al., 2001). Based on the calculated descriptor and feature extraction algorithm implemented, a matching algorithm calculates a similarity score that indicates the visual similarity level between images, and the classification algorithm decides if the input image belongs to the required reference classes or not. The resulted scores along with the IDs of the compared images or, in the case of the classification function, the decision whether the input image belongs to the inquired class or not, are sent to the Knowledge Repository as an Extensible Markup Language (XML) file that will be available to the system.

In this way users can detect required images not only among other images but also broaden the results looking inside the videos crawled by the system as additional information.

3.4 Face recognition on video frames

In the context of the discussed solution, biometrics is used to search for predefined subjects of interest within the processed videos. Face biometrics technology is utilized in this task as it represents the most clearly available biometric characteristic in such a scenario, as well as being one of the most thoroughly studied and widely accepted biometric solutions. A biometric solution consists of two processes, the enrolment and the recognition. Enrolment defines a reference model for the subject of interest where the recognition process identifies or verifies the identity of the captured subject. In the following, both processes are discussed.

For enrolment, a number of face images are used to create a biometric reference for the subject of interest. Those images can be different frames of a video sequence or static images of the same subject. The main face is detected in each image using multi-scale sliding window detection as proposed by Viola and Jones (Viola and Jones, 2001). The detected faces are normalized then passed on to the pose alignment

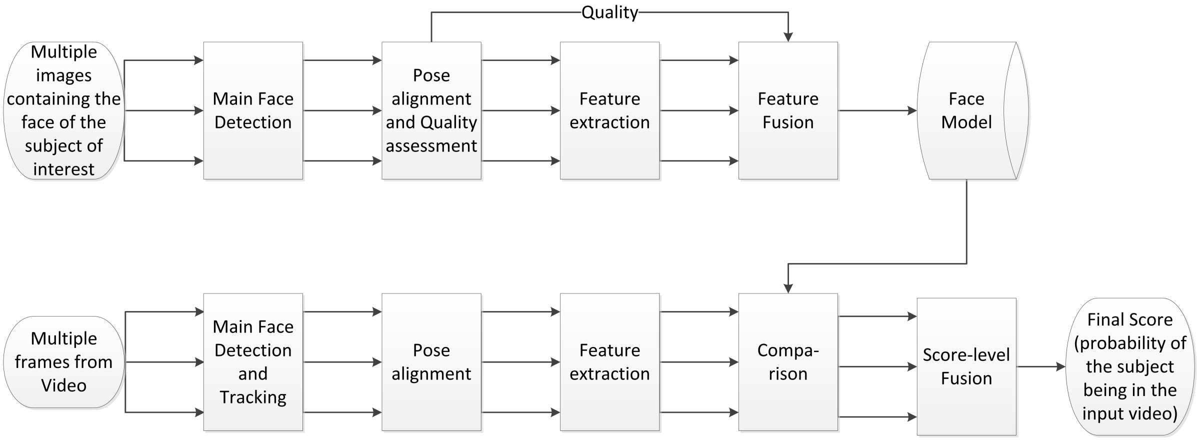
Quality
Multiple images containing the face of the subject of interest
Main Face Detection
Pose alignment and Quality assessment
Feature extraction
Feature Fusion
Face Model
Multiple frames from Video
Main Face Detection and Tracking
Pose alignment
Feature extraction
Compa-rison
Score-level Fusion
Final Score (probability of the subject being in the input video)

Figure 3: Overview of the proposed face recognition in video solution.

module. Pose alignment here is based on the unsupervised joint alignment of complex images presented by Huang et. al. (Huang et al., 2007). A binary feature vector based on the Local Binary Linear Discriminant Analysis (LBLDA) (Fratric and Ribaric, 2011) is extracted from each face image. Those feature vectors are then fused to create a robust feature vector that represents the subject of interest.

To search for the subject of interest in a video sequence, faces are detected and tracked across the frames (when possible). Those faces are normalized, the face pose is aligned and then passed on to feature extraction. Each feature vector (resulted from one face image) is compared to the reference face model resulting in a similarity score value. The score values related to the tracked face in different frames (if available) are fused using simple combination score-level fusion rules (Damer et al., 2013). The resulting fused score indicates the probability that the processed video (or image) contains the face of the subject of interest. An overview on the proposed biometric solution is presented in Figure 3.

3.5 Speaker identification

In the workflow of the proposed solution, speaker identification is performed on the audio crawled or extracted from a video and saved in the shared repository.

The algorithm creates models for a set of speakers associated to the specific RL and searches for those speakers in all audio files processed by the system. It is responsibility of the system operator to define the speakers of interest and to request the creation of speaker models to the system. When an audio is processed, the algorithm loads the current speaker models, provides audio analysis and outputs an XML file per audio file with the speaker segments belonging to the speaker models.

The algorithm works in two steps: creating models for specific speakers and audio analysis to identify the speakers previously defined by models. The algorithm is based on a Bayesian Information Criterion (BIC) segmentation followed by a BIC clustering. It starts by detecting speaker turns using BIC, where change points are detected through generalized likelihood ratio (GLR), using Gaussians with full covariance matrices. Speech-Non-Speech (SNS) segments are also modeled with Full Gaussian and compared with the current speaker Full Gaussian. If the BIC score is lower than 0, then a merge is performed between the current SNS Full Gaussian and the current speaker Full Gaussian. If the BIC score is higher than 0, then a new speaker Full Gaussian is created and a hierarchical clustering is performed for the previous speaker cluster. In our hierarchical clustering algorithm, the current speaker cluster, provided by turn detection and modeled with Full Gaussian, is compared with the clusters obtained so far. This method allows on-line processing of the clusters.

The speaker identification component works after the speaker clustering. Our speaker identification component uses the low-dimensionality total variability factors (i-vector) produced by the Total Variability technique to model known speaker identities (Dehak et al., 2011). Since this component works on-line, every time an unseen speaker starts talking, the component is incapable of knowing the speaker identity immediately. To overcome this problem, a first estimate for his/her identity is produced(if he/she is a known speaker) after 10 seconds of speech and a final identity estimation after 30 seconds. Since the zero and first-order sufficient statistics (and the respective i-vectors) from Total Variability are associated with

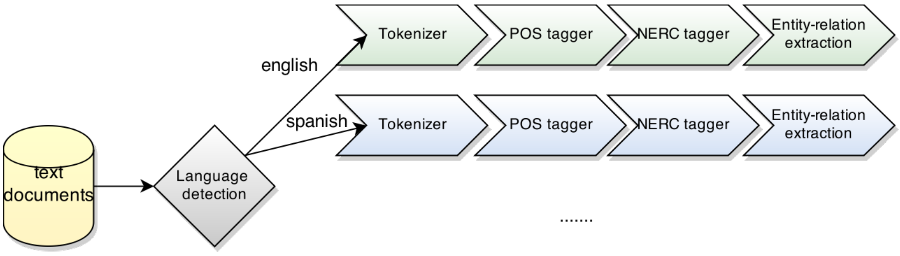

Figure 4: Overview of text analysis architecture

the cluster, the speaker information is immediately available whenever a cluster with a known identity appears.

As output of the algorithm an XML file is generated with the timing information of all speaker clusters and the IDs associated to the segments whose speakers are registered and of which the system has models.

3.6 Text analysis: entity and entity relationship recognition

In the workflow of the proposed solution, text analysis is performed on the audio extracted and transcribed from videos and saved in the shared repository.

The text analysis module is the responsible for the analysis in different languages of any input given in the project. The languages considered are Catalan, Spanish, English, French, Italian, Portugues, German, Romanian, Basque, Chinese, Japanese, Hebrew, Arabic and Russian.

The workflow used in particular for the last part of the workflow chain focused on the text analysis contains 4 different core components. The first component named language detection detects the language of the document. The second component named Part-of-Speech (POS) tagger analyzes the morphosynthactic category of each word in a document. The third component named Named Entity Recognition and Classification (NERC) (Nadeau and Sekine, 2007) detects and categorizes the Named Entities (NE) contained in a document. The main clustered categories are Person, Location, and Organization. Finally, the fourth and last component uses the extracted NEs to give an output based on the relationship (weighted-distance based) between the different entities in the whole document. This last component is used in the Visual Analytics component for visualization purposes.

Text analysis is performed in a webservice based architecture where the different language processors are host in different servers.

Figure 4 shows the general schema of the Text Analysis workflow. The language identification component detects the document language and sends it to the language processor for that language.

NE networks as the result of text analysis are encoded in a layered XML-based format named Knowledge Annotation Format (KAF) (Bosma et al., 2009), which is saved in the shared repository once the processing has finished.

4 CONCLUSIONS AND FUTURE WORK

The increasing need of LEAs for the monitoring and analysis of open sources boosts the implementation of systems capable of extracting relevant information from multimodal data. This work presents an effective platform for the analysis of all these crawled data and an intuitive results presentation that eases the cues searching process. The high implication of European LEAs has bolstered the development of the required solutions leading the presented system to the achievement of its goals.

Validation of this OSINT system has not been undertaken yet. The system final version's integration is currently ongoing and its validation by LEA end-users will take place by the end of the project.

In order to validate the results of the system a test scenario has been defined. LEA end users will run the test in two parallel researches, one using their existing tools and another one using the new OSINT system. Several metrics have been defined to assess not only the quantity and quality of the results obtained by each method, but also the time spent (human effort) and the total duration of the process.

As aforementioned, the validation process is still pending, but the undertaken tests have so far highlighted the necessity of approaches that nourish the system with deeper knowledge about the gathered information. This knowledge is foreseen to be extracted by the implementation of new analysis modules connected to the CMA. Still more, the integration of advanced visual analytics algorithms for the results presentation interface will tackle the current situation where a huge amount of data does not directly imply a huge amount of insight.

REFERENCES

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). Surf: Speeded up robust features. In *In ECCV*, pages 404–417.

Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*.

Chen, Y., Zhou, X., and Huang, T. S. (2001). One-class svm for learning in image retrieval. pages 34–37.

Damer, N., Opel, A., and Shahverdyan, A. (2013). An overview on multi-biometric score-level fusion - verification and identification. In Marsico, M. D. and Fred, A. L. N., editors, *ICPRAM*, pages 647–653. SciTePress.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798.

Ejaz, N., Tariq, T. B., and Baik, S. W. (2012). Adaptive key frame extraction for video summarization using an aggregation mechanism. *Journal of Visual Communication and Image Representation*, 23(7):1031–1040.

Fratric, I. and Ribaric, S. (2011). Local binary lda for face recognition. In *Proceedings of the COST 2101 European conference on Biometrics and ID management*, BioID'11, pages 144–155, Berlin, Heidelberg. Springer-Verlag.

Huang, G. B., Jain, V., and Learned-Miller, E. G. (2007). Unsupervised joint alignment of complex images. In *ICCV*, pages 1–8. IEEE.

Lowe, D. G. (2003). Distinctive image features from scale-invariant keypoints.

Maybury, M. T. (2012). *Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance and Authoring*. Wiley.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Rui, Y., Huang, T., and Mehrotra, S. (1998). Exploring video structure beyond the shots. In *Multimedia Computing and Systems, 1998. Proceedings. IEEE International Conference on*, pages 237–240.

Sun, Z. and Fu, P. (2003). Combination of color- and object-outline-based method in video segmentation.

Tuytelaars, T. and Mikolajczyk, K. (2008). K.: Local invariant feature detectors: A survey. *FnT Comp. Graphics and Vision*, pages 177–280.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511–I–518 vol.1.

Zhang, H., Kankanhalli, A., and Smoliar, S. W. (1993). Automatic partitioning of full-motion video. *Multimedia systems*, 1(1):10–28.

Zhang, J. F., Wei, Z. Q., Jiang, S. M., Li, J., Xu, S. J., and Wang, S. (2012). An improved algorithm of video shot boundary detection. *Advanced Materials Research*, 403:1258–1261.