VidTAG: A Learning Framework for Generating Action Tags in Videos

Naveed Ejaz, Sung Wook Baik*
naveed@sju.ac.kr, sbaik@sejong.ac.kr
Digital Contents Research Institute, Sejong University, Seoul, Republic of Korea

Abstract. *The volume of video media on the internet and from other sources is increasing at a rapid rate. The existing analysis, storage, retrieval, indexing and searching techniques are unable to cope with this huge volume of information. Moreover, there is no efficient way of extracting information from this huge pool. In this lieu, we propose a framework called 'Generating Action Tags and Videos' (VidTAG) for generating descriptions of videos. Our proposed framework for VidTAG is independent of domain specific techniques and can thus be applied for generating video descriptions in other domains. Video description can help in extracting useful information and patterns from the large volume of videos on the internet and improve existing indexing, searching and retrieval techniques. This paper describes the framework and the preliminary developments in implementation of the complete system.*

Keywords. Video Tagging, Action Recognition, Video Retrieval, Object Recognition

1 Introduction

With the widespread of the internet, advances in video and image compression technique, and cheap storage and capture mechanism the volume of image and video media is increasing at a tremendous rate[1]. According to a com Score report, from YouTube alone there are over 100 millions streams daily[2]. In such circumstances there is a need for efficient, robust and generalized media management tools. This management is not only limited to storage and retrieval but extends to effective meaningful search and automated association of semantics to the content[3]. Although the advances in computer vision have provided effective solutions for solving video data management problem, their scope is very limited. These include solutions such as automatic face tagging in Picassa[4]. Similarly Google's latest release Google Goggles is a great advance in content based image retrieval but it is also limited in scope.

The proposed system *'Generating Action Tags and Videos'* (VidTAG) aims to generate scene descriptions for simple videos using object and action classification in videos. Given a video, we propose a learning framework for generating video descriptions in a domain independent manner. We call these descriptions 'action tags'. VidTAG as a system can be broken down into two distinct functional phases; the learning phase and the classification phase. In the former, VidTAG will learn to classify objects and their actions in a supervised manner relying on a data set of tagged images and video. In the classification phase, VidTAG will classify objects and actions in input videos utilizing classification rules learnt in the former phase.

2 Scope and Objectives

As a matter of experimentation, the car and bike videos have been selected as the testing domain. Any analysis system developed for cars and bike videos has a ready application in traffic flow monitoring, road surveillance, automatic ticketing etc. This is one of the main reasons behind choosing car and bike videos as the domain. The problem of description generation for car and bike videos is not overly narrow and affords the avoidance of unnecessary complexity without the loss of generality. For instance actions performed by vehicles such as turning, passing by, jumping over a cliff etc are also common with a variety of other objects. So the mechanism for learning of classification rules can easily be extended to other objects.

In this context it must be clarified that the generality of the mechanism is not to be confused with the generality of the system. For instance an expert system always uses facts to infer rules and uses those rules to draw conclusion about given facts. Thus all expert systems follow the same mechanism. However, an expert system for finance cannot be used for geological surveys and a medical expert system cannot be used to fly an aircraft. Since a general mechanism does not imply a general system. Our objective in this project is to make a ground for the

*Corresponding Author

development of a framework that can be evolved into the basis of all expert systems in computer vision. Following are the main objectives of VidTAG.

- Learning object-models from a database of tagged images,
- Summarizing videos into representative key frames,
- Recognizing and tagging object-models in images,
- Extracting action-profiles for objects from videos,
- Learning action-profile tagging rules by extraction them from videos,
- Matching action-profiles of objects in a given video to generate action tags.

The main objective of VidTAG is to generate descriptions for simple car and bike videos. To properly define the scope of the project two things need to be defined; (1) what would be covered by the video descriptions, and (2) what is the exact meaning of simple videos. In VidTAG the video descriptions will cover simple short term actions of cars and bikes in the videos. Some examples of these actions can be 'a red car moving down a country road' or 'a blue car taking a right turn'. Moreover, we assume that the videos have already been cut down into shots, meaning that a single object performs a single action in a video.

3 Proposed Framework

The objective of VidTAG is to generate action tags for videos. It is designed to accomplish the main tasks. First task is to learn object-models and action-profiles and second is to tag object-models and action-profiles. For the generation of action tags in videos the proposed framework is shown in Figure 1. The detail of main components of the framework is now discussed.

3.1 Video Summarization Engine

This engine takes videos as input and summarizes them by extracting key frames. This is vital to the overall efficiency of the system as the bulk of video frames containing little information is removed and need not be processed. For the purpose of this engine the problem is defined as finding the minimal set of key frames that cover all significant events or maximize the number of key frames while minimizing redundancy of information in these key frames. A technique based on comparison of frames is developed for video summarization.

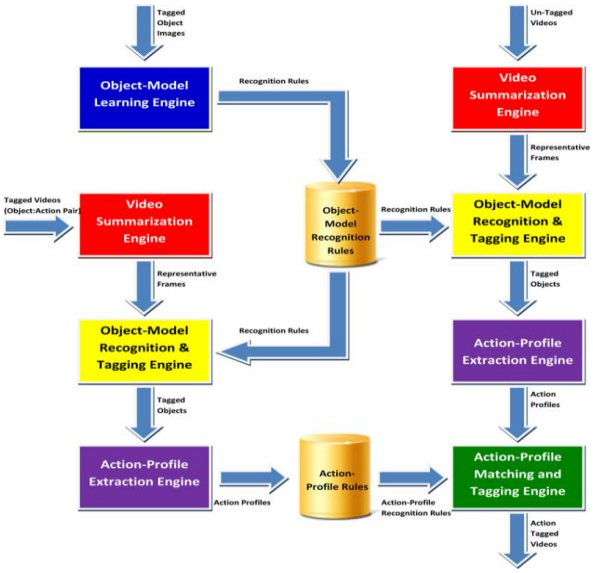

Figure 1: VidTAG Framework

The proposed technique compares the current frame with the last key frame instead of comparing consecutive frames. This allows for a summary that catches more significant events. Our utilized technique uses three comparison measures; correlation, color histogram difference and moment invariants. We pass the result of a comparison measure to the adaptive formula that combines them with the results of past comparisons. The result of the adaptive formula is compared with a threshold to obtain the confidence measure expressed by the said measure. A voting mechanism in used to combine the results of each measure. The details of our proposed technique for video summarization can be seen in [5].

3.2 Object-Model Learning, Recognition and Tagging Engine

3.2.1 Moving Object Segmentation

Segmentation of moving objects in video sequences is a vital part of VidTAG. It is a prerequisite for most object recognition techniques. In this section we discuss the moving object segmentation engine used by VidTAG. Given a frame in a video, we have to segment all moving objects in that frame. It is assumed that all objects of interest show movement in and around the given frame for segmentation. Another important assumption is that the video is shot from a static camera. This assumption is reasonable keeping in mind the overall scope of the project. MOSE uses a total of three consecutive frames for its operation; $frame_{i-1}$, $frame_i$ and $frame_{i+1}$, where $frame_i$ is the frame to be segmented. The technique first computes two different frames using the two pair of consecutive frames. These frames are then closed using morphological operations. Then the intersection of

these frames is computed, dilated and thresholded. This frame is combined with the edge detection output to compute the final output. This process is outlined in figure 2.

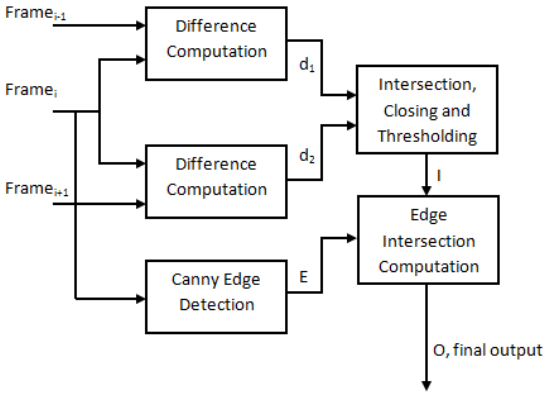

Figure 2: MOSE Framework

3.2.2 Object Recognition

This engine learns object-models from noise free tagged images during the learning phase and saves identification rules to an 'Object-Model Rules Database'. These rules are later used in identification of these objects. The Object-Model Rules Database stores the rules for recognition of object models. These rules are learnt using Machine Learning algorithms. Object-Model Recognition and Tagging Engine takes object-model recognition rules from the rules database to tags the key frames taken from the 'Video Summarization Engine'. These tagged objects are later used for matching action tags with action profiles. Object recognition is one of the most important problems in computer vision with wide ranging applications such as content based search, automated surveillance, action recognition etc. For this purpose, we used a framework for object recognition and pose estimation using SURF features with some modifications. Our feature-reduction process uses only the most repeatable features for matching. The noise-reduction process allows a further increase in matching speed-up reducing the false positive rates by 50%. A modified definition of the second-neighbor in the in the nearest neighbor ratio matching strategy allows matching with increased reliability. We also introduce a hierarchal approach for feature database storage that presents an easy way for pose estimation of objects. The details can be seen in [6].

3.2.3 Action Profile Extraction and Matching Engines

This engine uses the tagged key frames and object models to extract the action profiles of objects. During the learning phase these action profiles are combined with the provided action tags and saved in the 'Action Profiles Rules Database'. Action Profiles Rules Database stores the rules for recognition of action profiles of object actions. These rules are learnt using Machine Learning algorithms. This constitutes one half of the back bone of extensibility of the framework. Action Profile Matching Engine uses the output of the 'Action Profile Extraction Engine' and rules from the 'Action Profiles Rules Database' to tag actions of objects when an unseen video is provided to the system. These two engines are currently under development.

4 Progress, Evaluation and Bottlenecks

Thus far we have completed work in video summarization, moving object segmentation and object recognition. The results from these techniques are satisfactory. The assumption of a fixed camera for moving object segmentation, although reasonable, was not necessary for key frame extraction and object recognition. Thus this assumption was a bottleneck on the capabilities of the system. We removed this bottleneck by developing a technique for object recognition independent of image segmentation. Currently we are working on the Action Recognition engine. We have completed the literature review and are currently working on the framework for this component.

5 Preliminary Results

5.1 Results of Video Summarization Engine

For a detailed set of results of the proposed technique for video summarization, the reader are referred to [5]. In this paper, we show the results for a single sample video. The key frames extracted from the office tour video are shown above. There are two main reasons for testing on this video 1) It is easy to assess the quality of key frame extraction results 2) The lighting conditions vary greatly in the video allowing testing the system to the limits. The system is able to handle mild changes in lighting conditions (such as the bottom left frame). However, extreme changes in lighting conditions define (such as those in the third row) define an upper limit for insensitivity to lighting changes.

5.2 Moving Object Segmentation Results

This sub-section shoes the result of segmenting a moving object from video using the proposed moving object segmentation scheme. First three images in Figure 4 shows three consecutive frames used for

segmentation of 'moving car' by the proposed process. The rightmost image in Figure 4 shows the detected moving object.

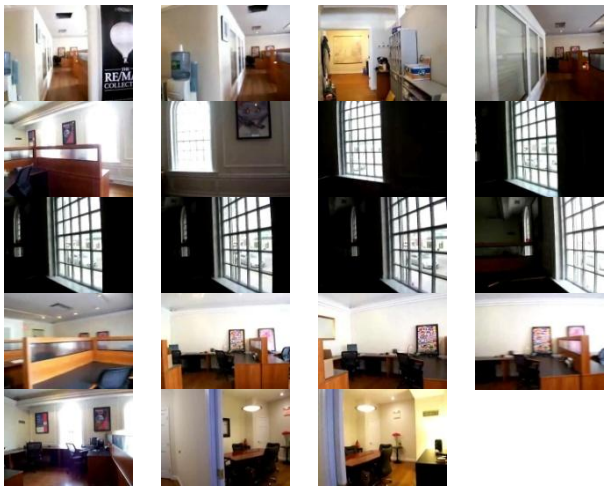

Figure 3: Key Frames extracted from Office Tour Video. Original Video Available at: http://www.youtube.com/watch?v=7qwp06xNuCs

Figure 4: Input frames for MOSE and output of moving object segmentation.

5.3 Object Recognition Results

The UK Benchmark Object Recognition Dataset was used for testing the object recognition module. The dataset contains 10,200 images of about 2500 objects. There are 4 images of each object. We used one image per object for training and three images per object for testing. With feature reduction a speed-up of 634.8% percent was achieved with less than 2% reduction in matching accuracy. After applying noise reduction on the reduced features the false positive rate was also pushed down by about 50%. After noise reduction an overall speed-up of 939.6% is achieved. For detailed results, please refer to [6].

6 Conclusions

VidTAG aims to generate descriptions for simple car and bike videos. We have not come across any projects in our literature survey that address the problem of video description. However, there are projects aimed at recognition of various actions and activities. Such projects can be regarded as addressing the problem of video description since video description and action recognition are closely related. The work is under progress on modules of action profile extraction and matching.

7 Acknowledgments

This work was supported by the Industrial Strategic technology development program, 10041772, The Development of an Adaptive Mixed-Reality Space based on Interactive Architecture) funded by the Ministry of Science, ICT & Future Planning(MSIP).

8 References

[1] Son J., Lee H., Oh H., “PVR: a novel PVR scheme for content protection”, IEEE Transactions on Consumer Electronics, 57(1): 173-177, 2011.

[2] "comScore Data Confirms Reports of 100 Million Worldwide Daily Video Streams from YouTube.com in July 2006," comScore, 2009.

[3] Huang Z., Li Y., Shao J., Shen H.T., Wang L., Zhang D., Zhou X., "Content-Based Video Search: is there a need, and is it possible?.", International Workshop on Information-Explosion and Next Generation Search, Shenyang, pp. 12-19, 2008.

[4] R. Broida, "Use Google Picasa to Face-Tag Your Photos," PC World, 2009.

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

[6] Ejaz N., Baik R., Baik S.W., "Feature Reduction and Noise Removal in SURF Framework for Efficient Object Recognition in Images", IT Convergence and Security, Lecture Notes in Electrical Engineering, 215: 529-535, 2013.

[7] Ejaz N., Baik S.W., “Video Summarization Using a Network of Radial basis Functions”, Multimedia Systems, 18(6): 483-497,2012.

[8] Ejaz N., Mehmood I., Baik S.W., "Efficient Visual Attention based Framework for Extracting key frames from videos", Signal Processing-Image Communication, 28(1): 34-44, 2013.