SHIATSU: Annotating Your Videos the Easy Way![*]

Ilaria Bartolini
DEIS, University of Bologna, Italy
i.bartolini@unibo.it

Corrado Romani
DEIS, University of Bologna, Italy
corrado.romani@unibo.it

ABSTRACT

In this demonstration we present SHIATSU, an automatic semantic-based video tagging system which relies on shot boundary detection and hierarchical annotation. More in details, in SHIATSU, shots obtained from video segmentation are first automatically labelled and such labels are then propagated at the video level. The approach is novel and appealing because: 1) it only considers the visual content of each video in order to automatically suggest description labels, 2) predicted tags can be reviewed/accepted by the user and persistently stored in the system in order to be exploited when searching for video of interest, and 3) the provided GUI makes annotation of videos extremely intuitive and usable.

1. INTRODUCTION

Searching for needed information from large video collections based on visual content and genre still represents a main open problem. This requires the definition of effective and efficient automatic video characterization and annotation techniques, so as to allow domestic/professional users to correctly retrieve videos of interest [4].

In this demo we present SHIATSU (Semantic HIerarchical Automatic Tagging of videos by Segmentation Using cuts), a semantic-based hierarchical video annotation system which first segments each video in scenes that are *similar* in term of visual features and then automatically provides compact descriptions by means of tags (at both shot and video level) which are useful to categorize the video content [2]. The annotation process exploits labels of pre-annotated key frames which are *similar* to the video to be labelled. A demonstration of SHIATSU on the TRECVID benchmark is provided by using the knowledge base supplied by the *TRECVID-2007 High-Level Feature* task [5].

[*]This work is partially supported by the CoOPERARE MIUR Project.

SISAP 2010, September 18-19, 2010, Istanbul, Turkey.

2. SYSTEM OVERVIEW

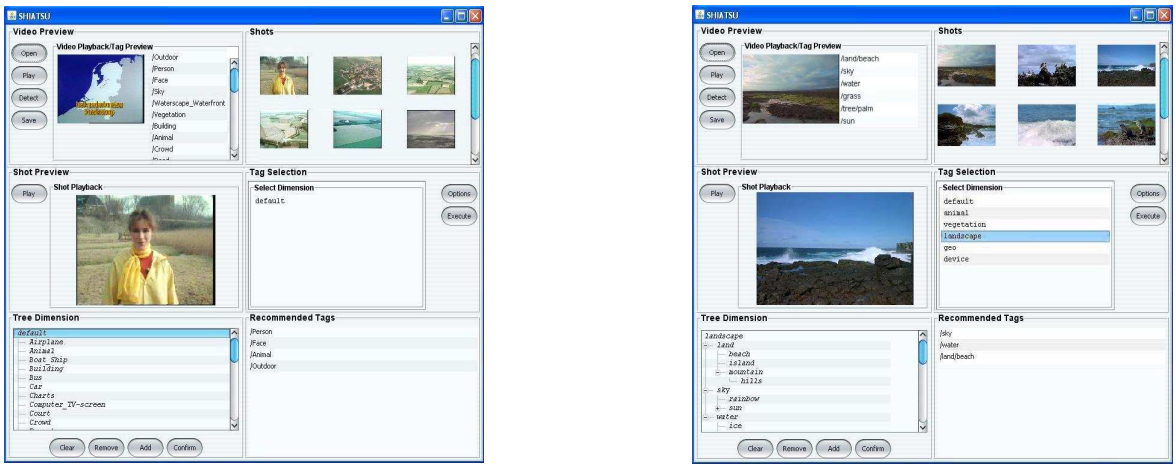
The core components of the SHIATSU system are a *video shot detector*, which fragments a video into coherent frame sequences, and a *video annotator* which attaches semantic concepts to such sequences. Shot tags are then propagated to the whole video, so as to obtain semantic indices for both the video and its shots: this allows the realization of a hierarchical, two-level, browsing platform.

When a video is processed, the shot detection component analyzes its frames and computes its shot boundaries, marking their timestamps (beginning and end of each shot). In details, color histograms and object edges are computed for every frame and such features are used to compare consecutive frames by applying two different distance metrics: this is because a shot transition usually produces a change in both the color and the texture structure of the frames. The shot selection process is done with a double dynamic threshold system which takes into account video content in order to adapt to different video types: frames are filtered on their color features and then on their edge features.

Every detected shot is then analyzed by the annotator component, which automatically assigns tags depending on the visual content of the shot. The user can then review the proposed tags and possibly modify the annotation results. After processing all the shots, the module selects the most appropriate tags for the whole video and saves all the information into a database. In details, the tagging phase exploits the Imagination system [1] and uses a set of pre-annotated images as a knowledge base. The system extracts a set of visual features from each image and saves the information in a database, indexing them efficiently with an implementation of the M-tree metric index [3].

The semantic concepts in the knowledge base can be organized either into a tree-shaped taxonomy named *dimension*, where terms are linked with a parent/child relationship, e.g., the term `landscape/land/beach` of the dimension `landscape` (see Figure 1 (b)) or as a flat structure, that we call `default` dimension, where all terms are at the same level (see Figure 1 (a)). Such dimensions can be easily modified and expanded.

When provided with a key frame to be labelled, SHIATSU first extracts its visual features. The key idea of tagging is to suggest those tags that are assigned to key frames in the knowledge base that are similar to the target frame (see [1] for more details). To efficiently perform similarity queries to key frames features, we exploit the M-tree index. Using the cuts timestamps, SHIATSU extracts a set of key frames representatives of each shot sequence, computes their vi-



(a) (b)

Figure 1: The SHIATSU interface: video tagging results for two videos when using the `default` dimension (a) and a specific dimension (i.e., `landscape`) (b).

sual features and compares them with those contained in the knowledge base. The annotator component suggests a set of concepts for each key frame: only terms recurring in the majority of key frames are selected as suitable concepts to describe the whole shot sequence. The number of key frames processed for every shot depends on the shot length. To avoid producing an overwhelming number of tags for each shot, only the most frequent tags retrieved for each key frame in the sequence are maintained (the total number of tags for each shot is also limited). The proposed tags can then be reviewed by the user.

Shot tags are useful to browse sequences across different videos, but they could be too specific to index a whole video, especially if this contains a wide range of different visual content. A simple criterion to select video tags from the set of shot tags is to weigh every tag depending on its frequency and the length of the shot it is associated to. Tags are ordered by descending values of weight and the first 10 tags (if available) become video tags. The rationale behind the proposed propagation method relies on the fact that concepts extracted from long shot sequences and/or that appear in several shots are probably more relevant, to describe the content of a whole video, than concepts occurring rarely or in short sequences. Finally, both shot and video tags are stored in the SHIATSU database, thus the user can exploit both of them when searching for her videos of interest.

3. DEMONSTRATION

Let us illustrate a possible usage scenario of SHIATSU (see Figure 1 for two real examples). First of all, the user opens a video by means of the "Open" button. The selected video can be played within the *Video Preview* frame by pressing the "Play" button, or segmented in shots trough the "Detect" option. In the latter case, the video is first processed by the video shot detector component and detected shots are then shown within the *Shots* frame.

At this point, the user can enable the tagging facilities by simply clicking on a specific shot. If this is the case, the user can play the shot within the *Shot Preview* frame by selecting the "Play" button. Then the user can select the dimension she prefers to consider during the annotation process; if user selection falls into the unstructured `default` dimension (SHIATSU default choice), as shown in the usage example of Figure 1 (a), all tags in the knowledge base are used as candidates for tagging. Coming back to our example, SHIATSU predicts tags `Person`, `Face`, `Animal`, and `Outdoor` which are, with exception for `Animal`, all relevant with respect to the shot content. On the other hand, if a specific dimension is selected (this is the case of the `landscape` dimension of the scenario depicted in Figure 1 (b)) the tags predicted by the annotator component will only contain concepts in that dimension. Continuing our running example, all predicted labels for the selected shot (i.e., `sky`, `water`, `land/beach`) are relevant.

Among tags predicted by the system for the shots, the user can refine them by deleting wrong labels and/or adding missing tags, depending on the precision of the provided results. When she is satisfied with the final result, she confirms it to the system which locally maintains provided annotations. Note that, in case a selected shot is already tagged, SHIATSU returns to the user the associated labels.

Any time a shot tagging is performed, SHIATSU propagates the labelling results at the video level by showing video tags within the *Video Preview* frame. Completing the running examples of Figure 1, video tags predicted by the system after annotating all the video shots are `Outdoor`, `Person`, `Face`, `Sky`, `Waterscape-Waterfront`, `Vegetation`, `Building`, `Animal` (not relevant), `Crowd` (not relevant), `Road` and `land/beach`, `sky`, `water`, `grass`, `tree/palm`, `sun` for the examples (a) and (b), respectively. At the end of the session, the user can save the result of the whole video annotation process for future use by selecting the "Save" button.

4. REFERENCES

[1] I. Bartolini and P. Ciaccia. Imagination: Accurate Image Annotation Using Link-analysis Techniques. In *AMR 2007*, pages 32–44, 2007.

[2] I. Bartolini, M. Patella, and C. Romani. SHIATSU: Semantic-Based Hierarchical Automatic Tagging of Videos by Segmentation using Cuts. In *AIEMPro 2010*. To appear.

[3] P. Ciaccia, M. Patella, and P. Zezula. M-tree: an Efficient Access Method for Similarity Search in Metric Spaces. In *VLDB 1997*, pages 426–435, 1997.

[4] P. Geetha, and V. Narayanan. A Survey of Content-based Video Retrieval. In *Journal of Computer Science*, 4(6), pages 474–486, 2008.

[5] TRECVID Video Retrieval Evaluation: http://trecvid.nist.gov.