Semantic Visual Templates: Linking Visual Features to Semantics

Shih-Fu Chang William Chen Hari Sundaram

Dept. of Electrical Engineering
Columbia University
New York New York 10027.
E-mail: {sfchang,bchen,sundaram}@ctr.columbia.edu

Abstract

The rapid growth of visual data over the last few years has lead to many schemes for retrieving such data. With content-based systems today, there exists a significant gap between the user's information needs and what the systems can deliver. We propose to bridge this gap, by introducing the novel idea of Semantic Visual Templates (SVT). Each template represents a personalized view of concepts (e.g. slalom, meetings, sunsets etc.), The SVT is represented using a set of successful queries, which are generated by a two-way interaction between the user and the system. We have developed algorithms that interact with the user and converge upon a small set of exemplar queries that maximize recall. SVT's emphasize intuitive models that allow for easy manipulation and queries to be composited. The resulting system performs well, for example with small number of queries in the "sunset" template, we are able to achieve 50% recall and 24% precision over a large unannotated database.

1. Introduction

The presence of vast digital repositories (both on-line and off-line) has lead to many schemes to index and retrieve such data. QBIC [Faloutsos 1993], VisualSEEk [Simth 1996] and Virage [Hamrapur 1997] are examples of a content-based paradigm as opposed to the old model based computer vision paradigm. Content based systems use feature vectors extracted from the data as indices for retrieval. Many systems also use keywords for retrieval when available.

The society of models approach [Picard 1996] uses multiple pre-clustered features in order to determine what combination of clusters best fits the user's query. It uses user feedback in order to determine the optimal partitions in the feature spaces. The feature vectors are derived by arbitrarily segmenting images into blocks rather than being derived more intiutively, using objects and scenes. The PicHunter [Cox 1996] system uses a Bayesian framework for image retrieval. The system does not allow for interactive querying but instead focuses on building a probabilistic model of the user's actions. However, despite some success, there are still many unresolved issues in content-based systems.

- In most feature based systems, the query procedure itself is non-intuitive making it extremely difficult for a naive user to use the system. The naive user is interested in querying the system at the semantic level rather than having to use features to describe his concept.
- Since the features by themselves are devoid of any semantics, a "good" match in terms of the feature metric may yield (and frequently does) poor results as far as the user is concerned.

Clearly, there is a significant gap between what the users desire and what content based retrieval systems can deliver today. In this paper, we propose to "fill-in" this link between semantics and low-level features with a new idea of *Semantic Visual Templates*. Semantic templates associate a set of exemplar queries with each semantic. The idea is that since a single successful query rarely completely represents the information that the user seeks, it is better to "cover" the concept using a set of successful queries. These queries are generated based on a two-way interaction between the user and the system. Each query in the template has been chosen since it has been successful at retrieving the concept that the user desires. The underlying feature set (and the query mechanism) can be general (global/local features; object based etc.). In this paper we demonstrate a proof of concept using VideoQ [Chang 1997] an object based retrieval system. There each query is in the form of a sketch.

1.1. Distinctive Features of SVT's

The generation of semantic template library involves no labeled ground truth data. What we do require is that the

some positive examples of the concept be present in the database. The following principles make the SVT distinct from traditional classification systems [Vailaya 1998].

Two-way Learning Our template generation system emphasizes the two-way learning between the human and the machine. We believe that since the human being is final arbiter of the "correctness" of the concept, it is essential to keep the user in the template generation loop. The user defines the video templates for a specific concept with the concept in mind. Using the returned results and relevance feedback the user and the system converge on a small set of queries that "best" match (i.e provide maximal recall) the user's concept.

Intuitive Models We use intuitive, understandable models for semantic concepts in the videos. The final sets of SVT's can be easily viewed by the user. Users can have direct access and make manipulation to any template in the library.

Synthesizing new Concepts Different SVT's can be graphically combined to synthesize more complex visual templates. For example, templates for "high-jumpers" and "crowds" can be combined to form a new template for "high-jumpers in front of a crowd."

Using the idea of semantic templates requires the two key components: An automatic generation mechanism and a metric that matches our sense of perception. We begin our description of SVT's by briefly reviewing VideoQ.

2. Background: VideoQ

The idea of the SVT is a general one, but we use the VideoQ[1]. framework in order to prove the concept. There, video streams undergo the following preprocessing steps:

1. The first step in the process is automatic scene cut detection [Chang 1997].

2. Automatic object segmentation and tracking after camera motion modeling and compensation.

3. For each object that has been segmented out, we extract attributes such as color, texture, size, shape and motion from the video shot. Additional information on the spatio-temporal arrangements of the video objects is also extracted.

The video object database consists of the all the objects extracted in the region segmentation process. Note that accurate segmentation (which in itself is a hard problem) of the real world video objects is not necessary for good results in video indexing. Rather, the objective is to extract and index the attributes of the salient objects in the video. It is the rich set of visual attributes and spatio-temporal relationships amongst the objects that turn out to be critical for retrieval. Figure 1 shows a query to retrieve downhill skiers (the user uses a sketch based query interface using a Java applet).

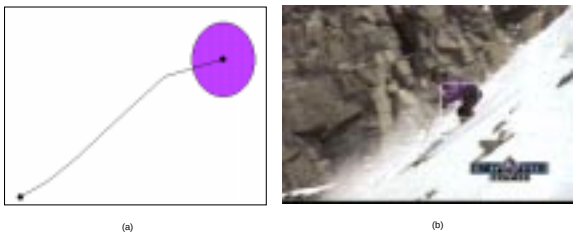


Figure 1. Visual query (a) and the corresponding search result (b) in the VideoQ system. The highlighted box shows the object matched to the object in the query. Video courtesy of Action Sports Adventure Inc.

3. Semantic Visual Templates

A visual template is a set of icons or example scenes/objects (feature vectors are extracted from these example scenes/objects for the query process) that represent the semantic that the template is associated with. The icons are animated sketches used in VideoQ. There, the features associated with each object and their spatial and temporal relationships are important. A typical example of a global features that could be part of a template are histograms, texture information and structural information. The choice between an icon based realization against a feature vector set created out of global characteristics depends upon the semantic that we wish to represent. For example, a "sunset" can be very well represented by using a couple of objects, while a waterfall or a crowd is better represented using a global feature set. Hence, each template contains multiple icons, example scenes/objects to represent the idea. The elements of the set can overlap in their coverage. The goal is to come up with a minimal template set with maximal coverage.

Each icon for the concept (e.g., down-hill ski, sunsets, beach crowds) is a visual representation comprised of graphic objects resembling the real-world objects in a scene. Each object is associated with a set of visual attributes (e.g., color, shape, texture, motion). The relevance of each attribute and each object to the concept is also specified. For example, for the concept "sunsets", color and spatial structures of the objects (sun and sky) are more relevant. The object "sun" may be non-mandatory since some

[1] http://www.ctr.columbia.edu/VideoQ/

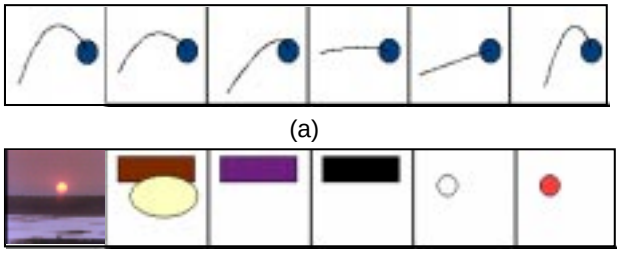


Figure 2. Semantic visual templates for the concepts (a) "high-jumper" Note that the color variation is not apparent (b) "sunset" The first icon is an example scene

sunset videos may not have the sun visible. For the concept "high jumper", the motion attribute of the foreground object (mandatory) and the texture attribute of the background object (non-mandatory) are more relevant then other attributes. Some concepts may need just need one object to represent the global attributes of the scene. Figure 2 shows several potential icons for sunsets and high-jumpers. The optimal set of icons would be chosen based on relevance feedback and maximal coverage in terms of recall. The generation of the icon set is explained in greater detail in Section 5.1.

4. Template Metric

The fundamental video data unit in our system is the video shot, which comprises multiple segmented video objects. The lifetime of an individual video object may be equal to or less than the lifetime of the entire video shot. The similarity measure D between a member of the SVT set and a video shot is the defined as follows:

$$D = \min \left\{ \omega_f \cdot \sum_i d_f(O_i, O_i') + \omega_s \cdot d_s \right\} \qquad (1)$$

where O_i are the objects specified in the template, O_i' are the matched objects for O_i, d_f is the feature distance between object O_i and O_i', d_s is the similarity between the spatial-temporal structure in the template and that among matched objects in the video shot, ω_f and ω_s are the normalizing weights for the feature distance and the structure dissimilarity. The query procedure that we adopt involves generating a candidate list for each object in the query. Then, the distance D is the minimum over all possible sets of matched objects that satisfy the spatio-temporal restrictions. For example, if the semantic template has three objects and two candidate objects are kept for each single object query, there will be (a maximum of) eight potential candidate sets of objects considered in computing the minimal distance in Equation 1. Note, that given N objects in the query, one could also exhaustively search against all set of N objects that appear together in one video shot. This however, proves computationally very expensive. This integrated feature-structure matching technique has been additionally successfully tested on our video search engine VideoQ. For querying the database using SVT's we use the following matching procedure:

1. Each video object (say O_i) is used to query the entire object database. A short list of matched objects is obtained by rank thresholding the rank list for the query using object O_i. Objects included in the shortened list are considered as candidate objects matching object O_i.

2. We then proceed to join the candidate objects on the short lists, thus obtaining the final set of matched objects, on which the spatial-temporal structure relationships will be verified.

5. Template Generation

Automatic generation of the SVT's is a hard problem. Hence we use a two-way interaction between the user and system in order to generate the templates. In our method, given the intial query scenario and using relevance feedback, the system converges on a small set of icons that gives us maximum recall. We now explain in detail the mechanism for generation of SVT's.

5.1. Generating Visual Templates

The user comes to the system and sketches out the concept that for which he wishes to generate a template. The sketch consists of several objects with spatial and temporal constraints. The user can also specify if the object is mandatory or not. Each object is comprised of several features. The user also assigns relevance weights to each feature of each object. This is the initial query scenario that the user provides to the system.

The initial query can also be viewed as a point in a high dimensional feature space. Clearly, we can also map all videos in the database in this feature space. Now, in order to generate the test icon set automatically, we need to make jumps in each of the features for each object. Before we do so, we must first determine the jump step size i.e. quantize the space. This we do with the help of the weight that the user has input along with the initial query. This weight can be thought of as the users belief in the relevance of the feature with respect to the object to which it is attached. Hence,

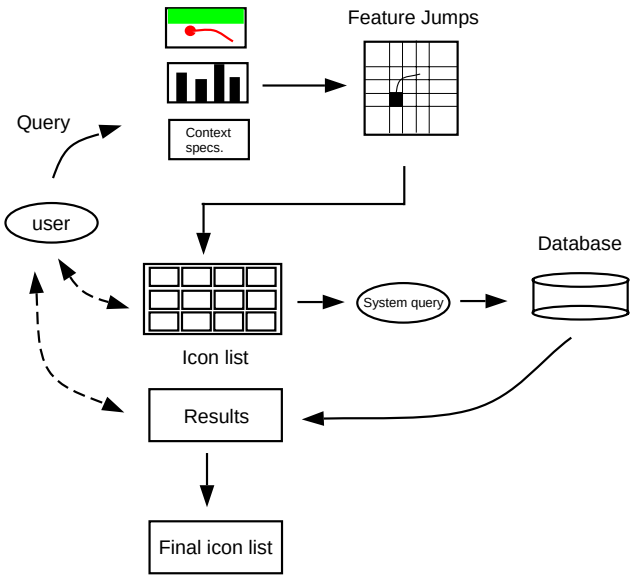


Figure 3. Interactive Template Generation. The dashed lines indicate a two-way interaction.

a low weight gives rise to coarse quantization of the feature and vice versa. In particular, we use:

$$\Delta(\omega) = \frac{1}{a \cdot \omega + b}, \tag{2}$$

where Δ is the jump distance in the metric space corresponding to that feature, ω is the weight associated with that feature, and a and b are parameters such that: $\Delta(0) = 1; \Delta(1) = d_o$. Using this jump distance, we quantize each feature space into hyper-rectangles. For example, for color, we generate the cuboids using the metric for the LUV space along with $\Delta(\omega)$.

Since the total number icons possible using this technique increases very rapidly, we don't allow for joint variation of the features. For each object in the query we do the following:

1. For each feature in the object, the user picks a plausible set for that feature.

2. The system then performs a join on the set of features associated with the object.

3. The user then picks the joins that most likely represent variations of the object. This results in a candidate icon list.

In a multiple object case, we do an additional join with respect to the candidate lists for each object. Now, as before the user picks the plausible scenarios. After we have generated a list of plausible scenarios we then query the system using the icons the user has picked. Using relevance feedback on the returned results (the user labels the returned results as positive or negative), we then determine the icons that provide us with maximum recall.

5.2. A Detailed Example

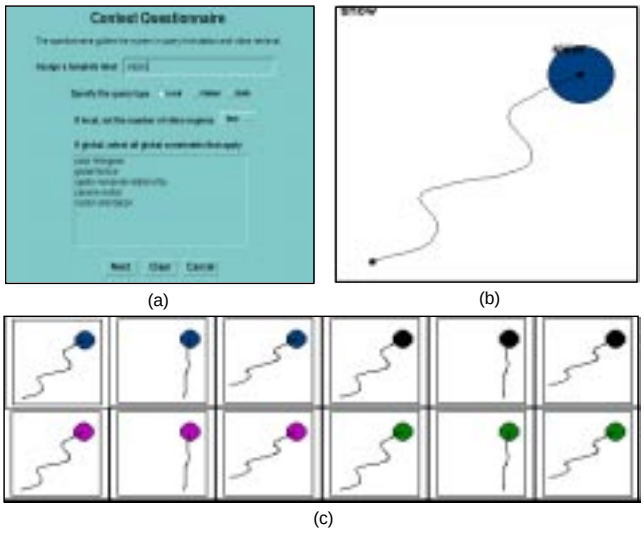


Figure 4. A Slalom template generation example

We generate semantic visual templates in order to retrieve video shots of slalom skiers.

1. We begin the procedure by answering the context questionnaire shown in Figure 4 (a). We label the semantic visual template "slalom". We specify that the query is object-based and will be composed of two objects.

2. In Figure 4 (b), we sketch the query. The large, white background object is the ski slope and the smaller foreground object is the skier with its characteristic zigzag motion trail.

3. We assign maximum relevance weights to all the features associated with the background and skier. We also specify that the features belonging to the background will remain static while those of the skier can vary during template generation.

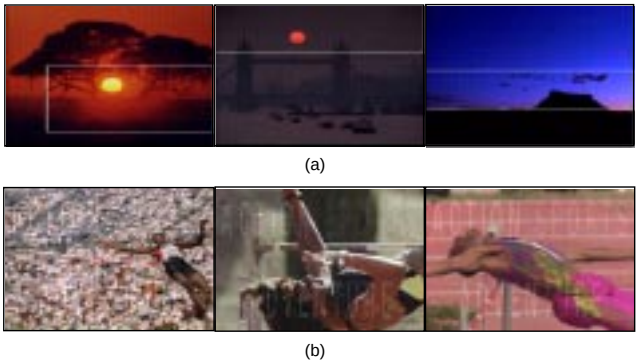
4. The system automatically generates a set of test icons, and we select plausible feature variations in the skier's color and motion trajectory.

5. The four selected colors and the three selected motion trails that belong to the skier are joined to form 12 possible skiers. The list of skiers is joined with the background. Note that the white background is static

throughout so the 12 skiers are joined with a single white background to generate the 12 icons shown in Figure 4 (c).

6. The user choses a candidate set to query the system. The 20 closest video shots are retrieved for each query. The user provides relevance feedback, which guides the system to a small set of exemplar icons associated with slalom skiers.

6. Experimental Results



(a)

(b)

Figure 5. The experimental results

We tested semantic visual templates for two classes of video shots: sunsets and high jumpers.

6.1. Sunsets

Our database contains 72 sunsets in over 1952 video shots. We first queried the system without semantic visual templates. Using just the initial visual sketch, we achieve 10% recall and 35% precision. We then generated eight sunset icons using semantic visual templates. With eight icons, we returned 36 sunsets, giving us a total recall of 50% and a precision of 24%. We note that the eight icons cover a wide variety of sunsets, as shown above.

6.2. High Jumpers

Our database contains nine high jumpers in 2589 video shots. Without semantic visual templates, we achieve 44% recall and 20% precision. With semantic visual templates, we achieve a slight improvement of 56% recall and 25% precision. We note that the system converges to a single icon, which is different from the initial sketch given by the user. Only a single icon is needed for two reasons. First we have a very limited dataset of high jumpers. Second, most of the high jumpers have very similar motions and camera views.

7. Conclusions

We have presented a new paradigm (Semantic Visual Templates) for bridging the gap between the low-level features that are derived from the raw data to semantics. SVT's are a personalized set of icons or example scenes/objects that characterize the concept well.

SVT's provides a mechanism for a two-way interaction between the user and the system. In our current scheme, the user provides the system with an intial skectch or an example image. This is the "initial seed" to the system to automatically genrate other "views" of the same concept. The user then picks the views that he thinks are plausible representations of the concept. Once the user and the system agree on a candidate list of views, the database is accessed and the user provides relevance feedback on the returned results. Then, the system determines the set that maximizes recall. The exprerimental results show that the idea of the SVT performs well leading to marked improvement in recall.

Once we have the visual templates for different concepts such as "skiing", "sunsets" etc. the user can interact with the system at the concept level. The user can compose a new concept by using these pre-existing library of templates. Clearly, the idea of using semantic templates is also easily generalizable to other media e.g audio.

References

[Chang 1997] S.F. Chang, W. Chen, J. Meng, H. Sundaram and D. Zhong *VideoQ: An Automated Content Based Video Search System Using Visual Cues*, ACM Multimedia 1997, pages 313-324, Seattle, WA. Nov. 1997.

[Cox 1996] I.J. Cox, M.L. Miller, S.M. Omohundro, P.Y. Yianilos *Target Testing and the PicHunter Bayesian Multimedia Retrieval System* ADL '96 Forum, Washington D.C. May 13-15 1996.

[Del Bimbo 97] A. Del Bimbo, P. Pala *Visual Image Retrieval by Elastic Matching of User Sketches*, IEEE Trans. on PAMI, Vol. 19, No. 2, pages 121-132, Feb. 1997.

[Faloutsos 1993] C. Faloutsos, M. Flickner, W. Niblack, D. Petkovic, W. Equitz, R. Barber, *Efficient and Effective Querying by Image Content*, Research Report #RJ 9203 (81511), IBM Almaden Research Center, San Jose, Aug. 1993.

[Hamrapur 1997] A. Hamrapur, A. Gupta, B. Horowitz, C.F. Shu, C. Fuller, J. Bach, M. Gorkani, R. Jain, *Virage Video Engine* SPIE Proceedings on Storage and Retrieval for Image and Video Databases V, pages 188-97, San Jose, Feb. 97.

[Picard 1996] R.W. Picard *A Society of Models for Video and Image Libraries* MIT Media Lab. TR #360, Apr. 1996.

[Simth 1996] J.R. Smith and S.F. Chang *VisualSEEK: A Fully Automated Content-Based Image Query System* ACM Multimedia 1996, Boston MA, Nov. 1996.

[Vailaya 1998] A. Vailaya, A. Jain, H.J. Zhang *On Image Classification: City vs. Landscape* in Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, Santa Barbara CA. Jun. 1998.