Efficient Co-Salient Video Object Detection Based on Preattentive Processing

Duan-Yu Chen and Chung-You Lin
Department of Electrical Engineering, Yuan-Ze University, Taiwan
dychen@saturn.yzu.edu.tw, s994633@mail.yzu.edu.tw

ABSTRACT
Automatic video annotation is a critical step for content-based video retrieval and browsing. Detecting the focus of interest such as co-occurring objects in video frames automatically can benefit the tedious manual labeling process. However, detecting the co-occurring objects that is visually salient in video sequences is a challenging task. In this paper, in order to detect co-salient video objects efficiently while maintaining the correctness, we first use the preattentive scheme to locate the co-saliency in a pair of video frames and then measure the similarity based on KL-divergence. Finally we improve the correctness of the matching across all video frames using our proposed filtering scheme. As a result, we are able to describe the significant parts of a video sequence based on the detection of co-occurring video objects. Our experiment results show that the proposed co-salient video objects modeling achieves high precision value about 85% and reveals its robustness and feasibility in video.

Keywords— Co-saliency, video object

1. INTRODUCTION
Methods for annotating video content as well as related video retrieval techniques have attracted a great deal of attention in recent years. Since video content contains much richer information than other types of media, the annotated content can be used in many fields, such as video surveillance systems and entertainment applications. Among the different types of video annotation approaches, detecting visual salience in a video is considered a good way to understand the video’s content. Moreover, detecting visual saliency successfully can substantially reduce the computational complexity of the video annotation process. According to a study conducted by cognitive psychologists [17], the human visual system picks salient features from a scene. Psychologists believe this process emphasizes the salient parts of a scene and, at the same time, disregards irrelevant information. However, this raises the question: What parts of a scene should be considered “salient”, especially for the scene in videos? To address this question, several visual saliency (or attention) models have been proposed in the last decade [1-5, 16, 18]. In videos, salient parts may depend on the contextual information in successive frames. Therefore, detecting the co-occurring objects across video frames would make the saliency process more specific. Therefore, in this work, co-salient video objects are the target that we aim to locate within each video clips.

In the literature [6-9, 10-11], the co-saliency problem in an image pair might be solved using some unsupervised algorithms to find the common parts among all possible correspondences. Previous approaches in the detection of co-salient objects generally would first extract the features in the regions of interest, e.g., SIFT [12], and then perform feature matching accordingly. However, the computation cost makes the process unpractical to be conducted in videos.

To speed up the process of the co-saliency computing in image pairs, Goldberger et al. [13] proposed two methods that approximats the KL-divergence between mixture densities. The first (match-based) method can be applied to any mixture density while the second (unscented) is tailored for mixtures of Gaussian densities. The efficiency and the performance of these methods were demonstrated on image retrieval tasks on a large database. In all the experiments conducted, the unscented approximation achieves the best results, results that are very close to large sample Monte-Carlo based ground truth. However, the KL-match based approximation is faster but less accurate than the unscented based method.

Therefore, in this paper, in order to detect co-salient video objects efficiently while maintaining the correctness, we first use the preattentive scheme [3, 15-16] to locate the co-saliency in a pair of video frames and then measure the similarity based on KL-divergence, and finally improve the correctness of the matching across all video frames using our proposed filtering scheme. As a result, we are able to describe the significant parts of a video sequence based on the detection of co-occurring video objects.

The remainder of this paper is organized as follows. In the next section, we introduce the proposed co-saliency modeling. Section 3 details the experimental results. Then, in Section 4, some concluding remarks are drawn.

2. CO-SALIENT VIDEO OBJECTS DETECTION
2.1. A Set of Preattentive Base Functions Used to Compute Salient Map

For detecting salient regions in video frames effectively, we shall have a representative set of base functions. Therefore, a set of sparse feature patches proposed by Hou and Zhang [15] is selected as the set of preattentive base functions that is of 192 8×8 image patches in RGB color space trained from a dataset of 120,000 images. Some patches are shown in Fig.1(c). The salient map for each video frame can then be obtained by applying the patch set. An example of salient map is demonstrated in Fig.1(b).

![Image](image.png)

Fig. 1. (a) an original video frame (b) The salient map of (a) obtained by finding the incremental coding length using the set of preattentive base function [15] (c) a part of the base functions.

2.2. Frame Response Corresponding to the Base Functions

Let \( B \) denote the matrix representation of the set of base functions, where each column of \( B \) corresponds to a patch of the base functions. With the set of base functions, a given image patch \( x \) can be represented by a linear combination of the base functions, such as \( x = B\alpha \). The coefficient set \( \alpha \) of the linear combination can be regarded as the response corresponding to each base function. Accordingly, \( \alpha \) can be obtained by \( \alpha = B^T x \).

To compute the response of a video frame, patches \( X = \{ x_1, x_2, \ldots, x_n \} \) are obtained by shifting a sliding window of size 8×8 in the video frame and then the normalized response \( p_i \) of each path is computed by

\[
P_i = \frac{\sum_{j \in w, p \neq q} F_i^T x_j}{\sum_k \sum_{j \in w, p \neq q} F_k^T x_j},
\]

where \( \sum p_i = 1 \), and \( w_p \) is the saliency weighting that is set 1 initially. \( F_i^T \) is the \( i \)th row vector of \( B \). The detail of weight updating scheme is described in the following section. \( \delta \) is a predefined threshold in the range \([0..1]\) that is used to filter out patches that are of relatively small response. The absolute value of the product \( F_i^T x_j \) is adopted since the magnitude of the response is the major concern. An example of the response using the set of base functions is demonstrated in Fig.1(b). It is clear that the highlighted area would generally be the focus of interest and is thus reasonably regarded as salient regions.

2.3. Response Similarity Estimation Using KL-Divergence

For a pair of frames \( X_S \) and \( X_T \), we then have a response probability distribution for each frame, say \( p \) and \( q \) and then compute the dissimilarity between \( X_S \) and \( X_T \) by KL-divergence defined as

\[
KL(p \parallel q) = \sum_i p_i \log \frac{p_i}{q_i}.
\]

(2)

Notice that we have to compute \( KL(p \parallel q) \) and \( KL(q \parallel p) \) respectively since these two values may not be the same and both \( KL(p \parallel q) \) and \( KL(q \parallel p) \) are all equal or greater than zero. In order to indicate which patches in base functions are more important to the minimization of the KL-divergence, the partial derivative of KL-divergence for \( X_S \) and \( X_T \) is computed respectively, i.e., for example, \( p_i \) for \( X_S \) is obtained by

\[
\frac{\partial}{\partial p_i} KL(p \parallel q) = \frac{\partial}{\partial p_i} \sum_j (p_i \log p_i - p_i \log q_i) = p_i \log p_i - p_i \log q_i - KL(p \parallel q) \]

(3)

Considering the characteristic of KL-divergence, the smaller KL-divergence would be ignored. For updating the patch weight, \( w_p \) is updated based on the refined KL-divergence and is defined as:

\[
KL'_p(p \parallel q) = \min \left( \frac{\partial}{\partial p_i} KL(p \parallel q), 0 \right).
\]

(4)

Eq.(4) means that patches with increase of the KL-divergence would be ignored. For updating the patch weight, \( w_p \) is updated based on the \( KL' \) and is defined by

\[
w_p = \frac{\sum_i KL'_p(p \parallel q)(F_i^T x_j)}{\sum_i KL'_p(p \parallel q)}.
\]

(5)

The updating process is terminated until the weight value is converged.

![Image](image.png)

Fig. 2. Demonstration of the detection process of co-salient video objects. (a)(d) original video frame pair (b)(c) co-salient maps (c)(f) detected co-saliency objects.

An example of co-saliency in a frame pair is illustrated in Figs.2(b)(e), where co-salient regions denote that the extent
is covered by some patches with small value of $KL'$. In Fig.2(b), co-occurring video objects can be addressed in co-salient map using the proposed approach. However, to find major co-salient objects across consecutive frames, a measure that can indicate the degree of similarity between co-salient objects is needed. The detail of this measure is described as follows.

2.4. Similarity Measure between Co-salient Objects
For co-saliency object detection, $KL'(p \| q)$ is used to select the patches that are co-salient in a frame pair. Considering a frame pair that is of certain co-saliency, $KL'(p \| q)$ would generally be similar to each other. Therefore, based on the difference of $KL'$ for each updating and the convergence rate, the similarity between co-salient objects can be estimated.

To compute the difference between co-salient objects, we select $\beta$ patches that are of the smaller $KL'(p \| q)$ and the saliency reflected by the $\beta$ patches is defined by

$$S_{KL} = \sum_{p_i \in \text{the } \beta \text{ smallest values of } KL}(1 + \frac{1}{e^{KL_{p_i}}}).$$

(6)

Accordingly the distance between two co-salient objects is defined as

$$D_{\beta}(p, q) = |S_{KL}(p) - S_{KL}(q)|.$$  

(7)

To model the difference of convergence rate between co-salient objects, the difference of $KL'$ in iteration $t$ and $t-1$ is modeled by

$$S_d = \sum_{t=1}^{\tau} \sum_{p_i} \left(KL'_{p_i(t)} - KL'_{p_i(t-1)}\right),$$

(8)

where $\tau$ is the number iterations. Notice that co-salient objects in a frame pair would have small value of $S_d$ if they are similar in appearance. An example of computing similarity between co-salient objects is illustrated in Fig.2(c) and Fig.2(f). As the case of Fig.2(c), it is clear that real co-salient objects can be addressed using our measure and the distinct ones shown in Fig.2(f) can be eliminated.

3. EXPERIMENTAL RESULTS
To evaluate the performance of the proposed visual saliency model, we conducted experiments on different kinds of videos of size $320 \times 240$ downloaded from Youtube. Some sample frames of the dataset are shown in Fig.3. In order to illustrate the qualitative evaluation of our proposed scheme, some detected co-salient objects are demonstrated. In Figs.3(b)(c), six most similar co-salient objects corresponding to source image are shown, i.e., from rank 1 to rank 6. The co-salient map for each frame pair is also shown next to its original frame to indicate where the co-salient objects are located. We can see that frames with similar co-salient objects corresponding to the source frame are detected. Particularly in Fig.3(c), co-salient objects, the roof of the hotel, of distinct size across different video frames can be correctly detected.
In this paper, in order to detect co-salient video objects efficiently while maintaining the correctness, we first use the preattentive scheme to locate the co-saliency in a pair of video frames and then measure the similarity based on KL-divergence. Finally we improve the correctness of the matching across all video frames based on our proposed filtering scheme. As a result, we can describe the significant parts of a video sequence based on the detection of co-occurring video objects. Our experiment results have shown that the proposed co-salient video objects model has achieved high precision value about 85% and reveals its robustness and feasibility in video.

### 4. CONCLUSION

In this paper, in order to detect co-salient video objects efficiently while maintaining the correctness, we first use the preattentive scheme to locate the co-saliency in a pair of video frames and then measure the similarity based on KL-divergence. Finally we improve the correctness of the matching across all video frames based on our proposed filtering scheme. As a result, we can describe the significant parts of a video sequence based on the detection of co-occurring video objects. Our experiment results have shown that the proposed co-salient video objects model has achieved high precision value about 85% and reveals its robustness and feasibility in video.

### 5. REFERENCES


