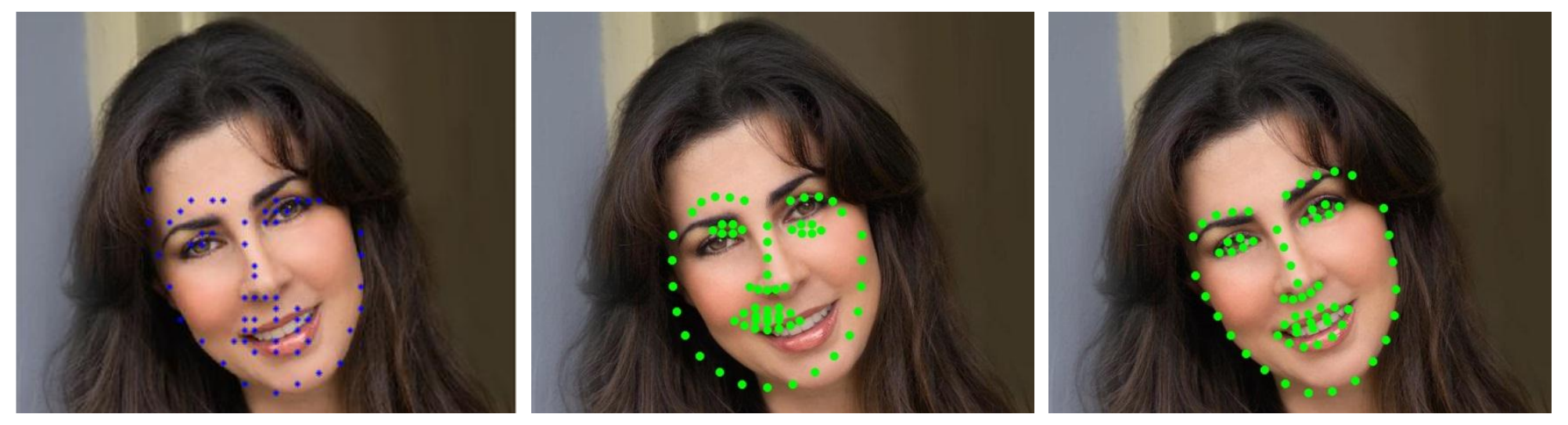
Approximate Structured Output Learning for Constrained Local Models with Application to Real-time Facial Feature Detection and Tracking on Low-power Devices

Shuai Zheng, Paul Sturgess, and Philip H. S. Torr

Oxford Brookes Vision Group, Department of Computing and Communication Technologies, Oxford Brookes University, Oxford, UK

OXFORD BROOKES UNIVERSITY

Introduction


Tree structured SVM

CLMs with BRIEF

Proposed

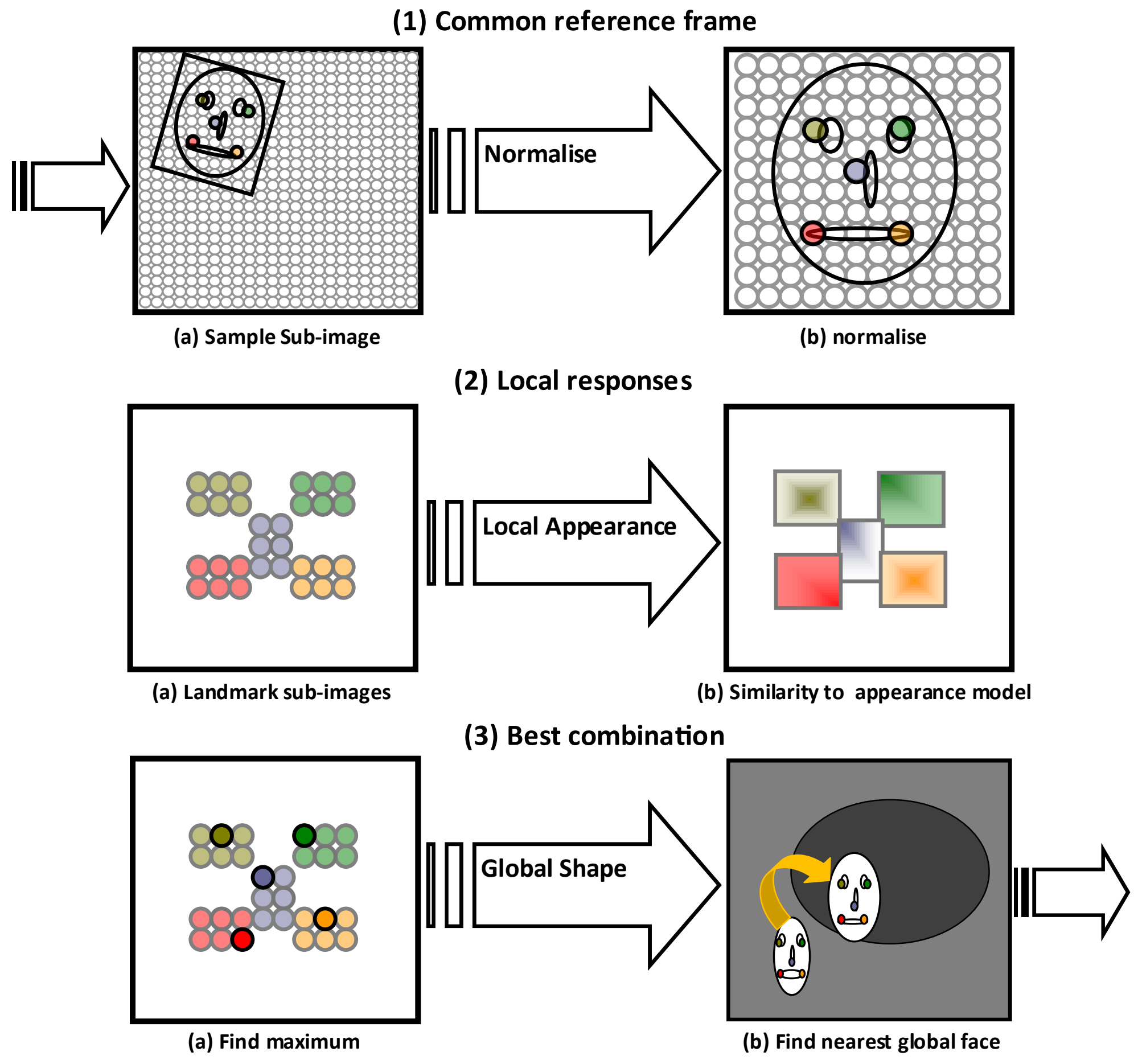
- Given a face detection, *facial feature detection* involves localizing facial landmarks.
- Constrained Local Models (CLMs) [4] are popular for facial feature detection as they are efficient.
- In this paper,inspired by the use of structured output support vector machines (SO-SVM) to learn pictorial structures in the object detection task, we propose adapting SO-SVM to learn the CLM appearance model parameters.
- Because the optimisation for CLM is not exact we explore approximate SO-SVM, interestingly we find that, although we lose the theoretical guarantees that approximate learning works well in practise.

Contribution

- We train the popular CLM approach by introducing approximate structured output optimisation to jointly learn the local appearance models.
- We approximate our CLM appearance model with a binary encoding scheme, which allows us to efficiently compute response maps with dot-product calculations, when using binary feature descriptors such as BRIEF.
- We demonstrate that our facial feature detector runs in real-time on low-power mobile device such as the popular ipad2 tablet, when integrated into a tracking-learning-detection (TLD) [3] system.

Inference for Constrained Local Models

- Apply a face detector [6] and then set the coordinates of the facial feature landmarks to the mean face of the CLM shape model.
- Generate a response map for each landmark by extracting the feature descriptors, and then computing a dot product between them and our facial feature classifier.
- Find the set of facial landmarks that best fit both the CLM appearance and shape models.


(1) Common reference frame

(a) Sample Sub-image

(b) normalise

(2) Local responses

(a) Landmark sub-images

(b) Similarity to appearance model

(3) Best combination

(a) Find maximum

(b) Find nearest global face

Learning for Constrained Local Models

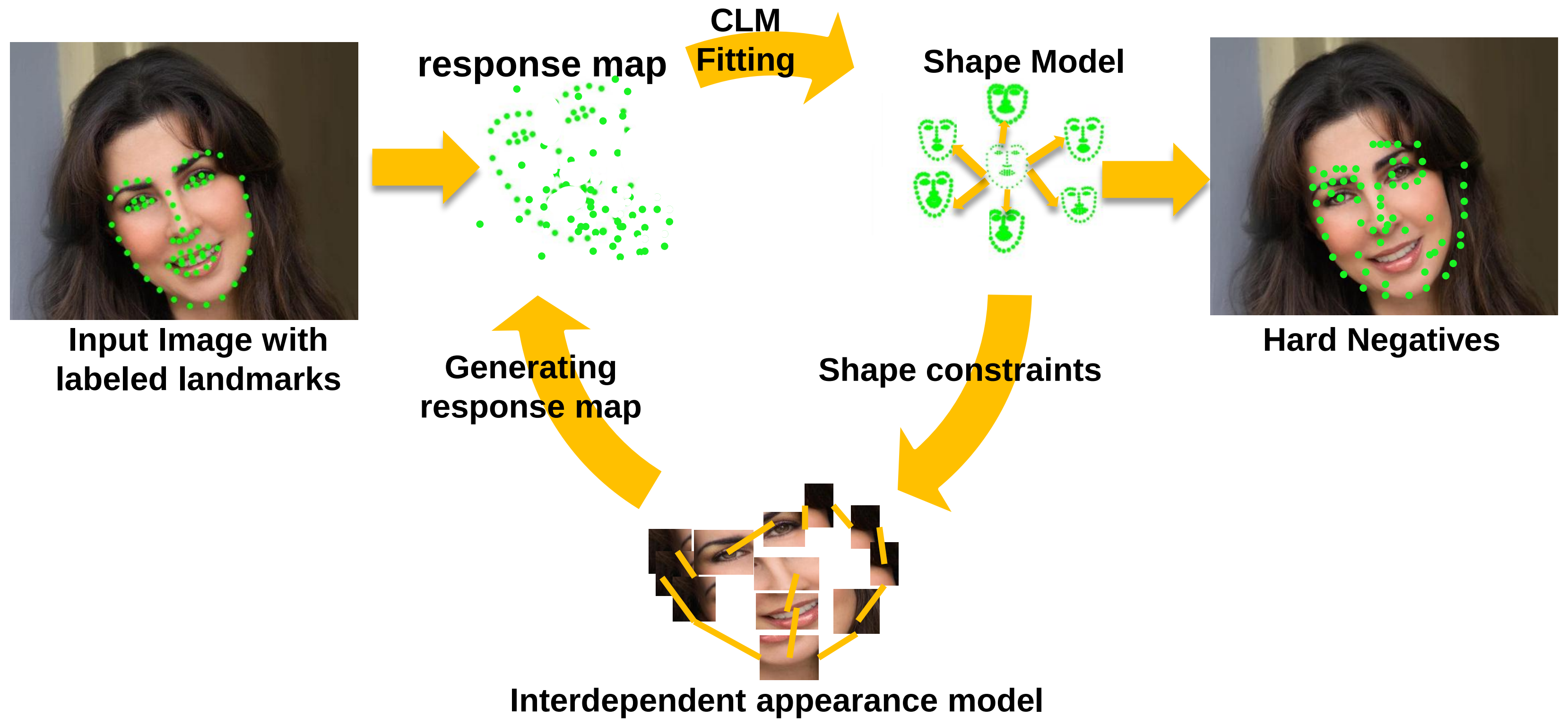
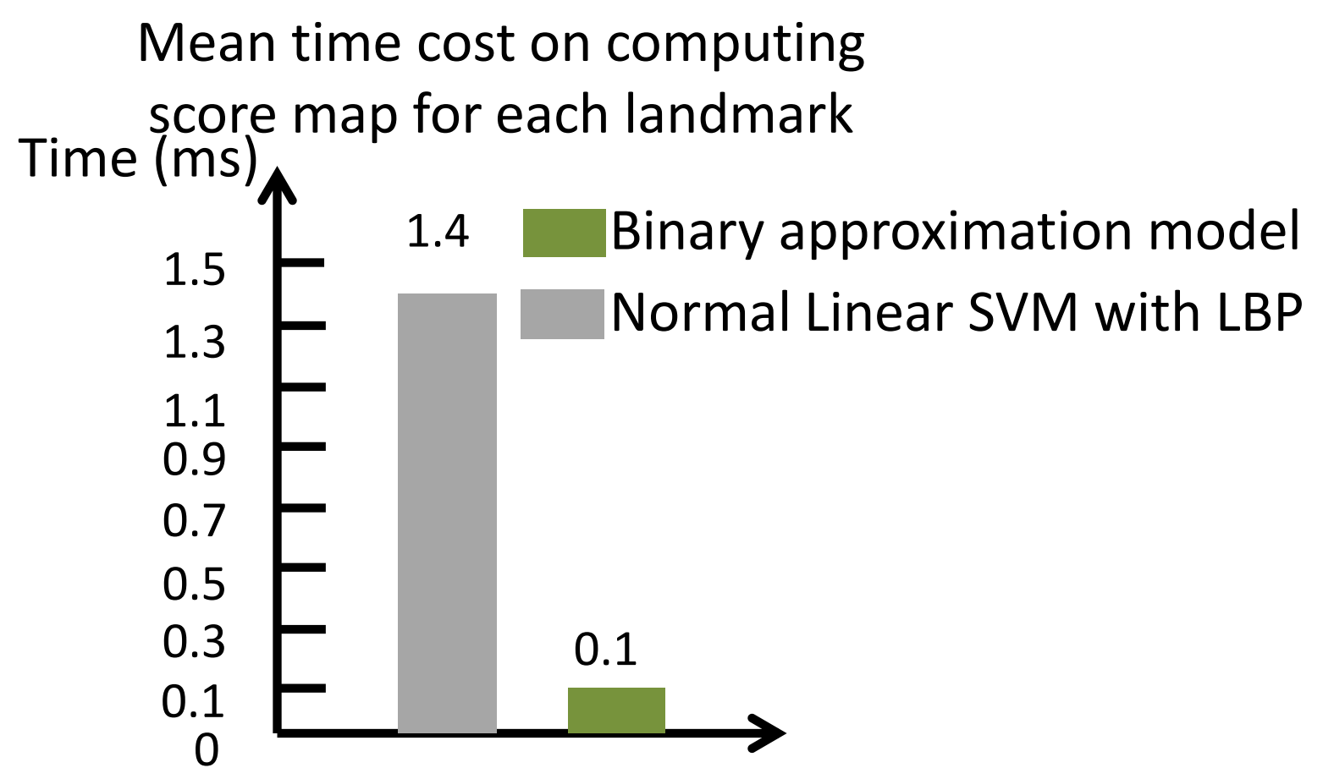


Figure: Approximate structured output learning for constrained local model.

- We learn the CLM shape model using PDM [1].
- We learn the CLM appearance model with approximate structured SVM:
  1. First initialise the CLM appearance model weights by applying a set of independent linear SVMs, one for each facial landmark.
  2. Given a set of true input-output pairs, find the nearest negative samples by employing CLM inference, we call these *hard negatives*.
  3. Given a set of true input-output pairs and the hard negatives, update the CLM appearance model weights with online SO-SVM [2].
- In step 2, the CLM search for hard negative is not guaranteed to find a global optimal, which is why we refer to our method as approximate structured learning.

Model Binary Approximation

- Following [2], we give an extra boost in speed by choosing to use a binary appearance features such as BRIEF, and a binary approximation of our learnt appearance model parameters, as shown in the figure.



Facial feature tracking-learning-detection

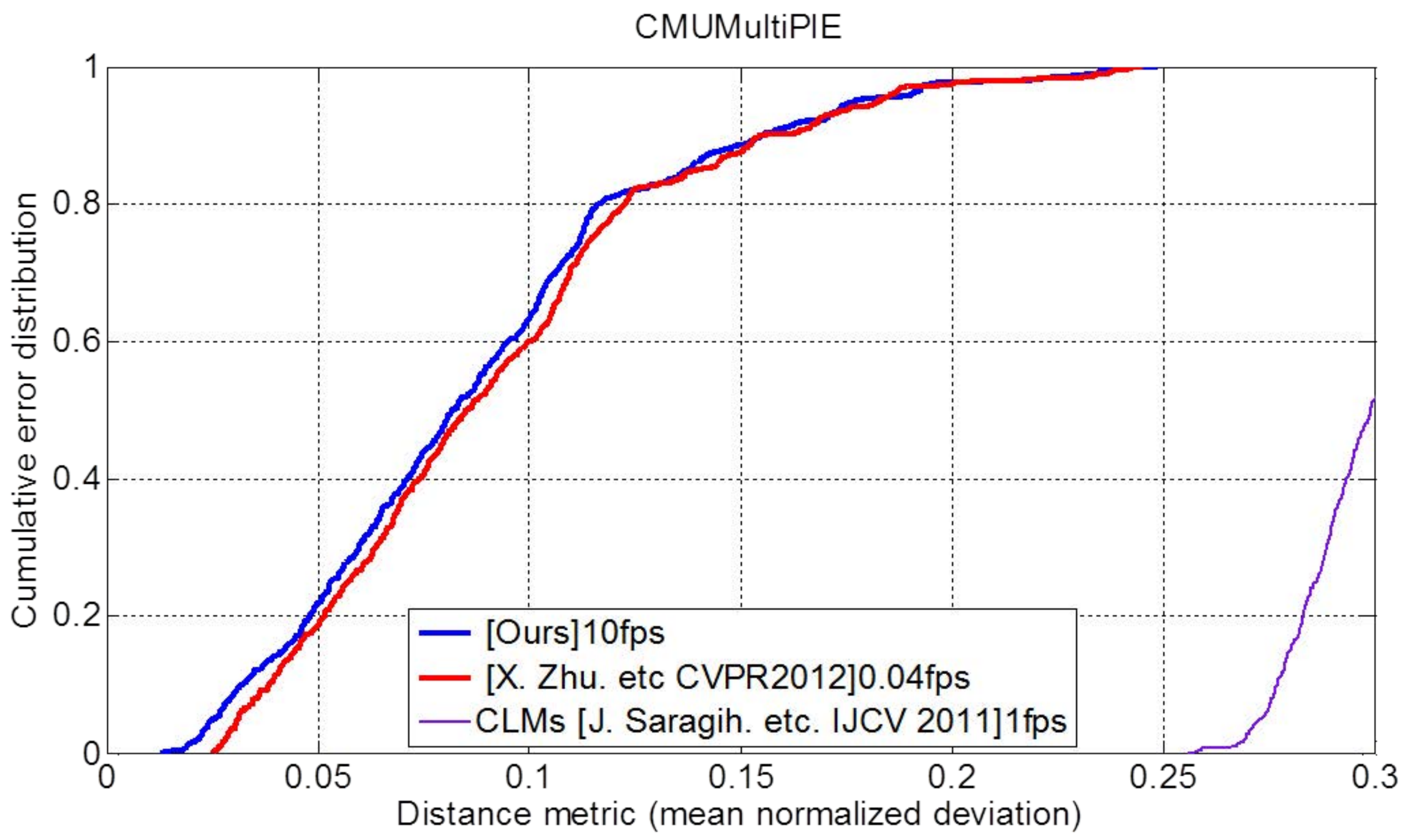
- We also develop a facial feature tracking system using the Tracking-Learning-Detection [3] scheme.
- TLD is a highly efficient method that meets our real-time requirements.
- TLD is dynamic, which means that we can increase accuracy over deploying our pre-learned model.
- We use the online SO-SVM [2] to perform appearance model updates.

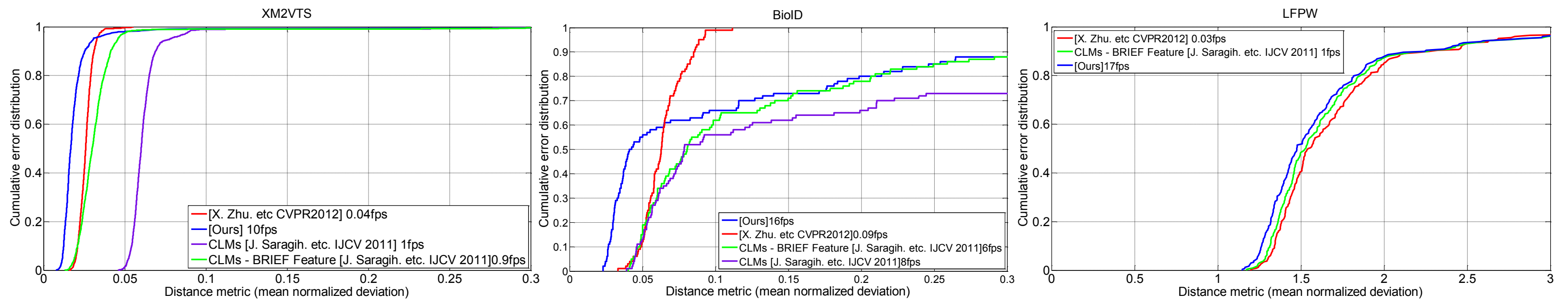
Acknowledgement

We would like to thanks EPSRC EP/I001107/1, and PASCAL2 Network of Excellence, IST-2007-216886.

Experiments

- From CMU Multi-PIE, we randomly choose 1225 images as training images, 734 images as testing images, and 491 images as validation images.
- To evaluate the generalisation of the proposed approach, we train a model on the CMU Multi-PIE and test on the XM2VTS, BioID, and LFPW datasets.
- We evaluate performance by computing the mean Euclidean distance for 17 estimated landmarks. This is then normalized with respect the Euclidean distance between eye centers in ground truth to render the mean invariant to scale [5].





- Performances and generalisation properties of our approach are comparable to the state-of-the-art [7].
- Our approach is more efficient at run-time compared to other approaches.

Figure: ipad2 Demo. This figure shows our SO-CLM running on an ipad2 low power device.

- Our approach runs in real-time on low-power devices such as the popular ipad2 tablet, and smart phones.

References

[1] T. F. Cootes and C. J. Taylor. Active shape models - 'smart snakes'. In *BMVC*, 1992.

[2] S. Hare, A. Saffari, and P. H. S. Torr. Efficient online structured output learning for keypoint-based object tracking. In *CVPR*, 2012.

[3] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pages 1–14, 2011.

[4] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting with a mixture of local experts. In *ICCV*, 2009.

[5] P. A. Tresadern, M. C. Ionita, and T. F. Cootes. Real-time facial feature tracking on a mobile device. *International Journal of Computer Vision*, pages 280–289, 2011.

[6] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[7] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.

http://cms.brookes.ac.uk/research/visiongroup/ szhengcvpr@gmail.com, paul.sturgess@brookes.ac.uk, philiptorr@brookes.ac.uk