FACE DETECTION BY FACETS:

COMBINED BOTTOM-UP AND TOP-DOWN SEARCH USING COMPOUND TEMPLATES

Glendon R. Holst
Department of Computer Science, University of British Columbia
gholst@cs.ubc.ca

ABSTRACT

New techniques are needed to effectively search the image and feature space of larger and more complex domains. One such technique uses subfeatures and spatial models to represent a compound object, such as a face. From these compound models, hypothesis based search then combines bottom-up and top-down search processes to localize the search within the image and feature space. Detected sub features become evidence for facial hypotheses, which then guide local searches for the remaining subfeatures, based upon the expected facial configuration. We describe this compound technique and present a comparison of the compound templates technique with a single template technique in a mug shot style face domain. Attention is paid to performance, including both efficiency and accuracy. The results are complex, and the strengths, weaknesses, and various trade-offs of the two techniques are detailed.

INTRODUCTION

Faces are diverse, semi-rigid, semi-flexible, culturally significant, and part of our individual identity. As a domain they are already well studied and many images databases already exist. For complexity the domain is extensible along a continuum ranging from expressionless frontal, to increasingly varied expression, to increasing pose variance, to increasing artistic interpretation.

The chosen domain is face detection of photographic, grey scale, near frontal, mildly expressioned, faces. This domain is practical yet interesting; challenging yet tractable. Examples of the domain are shown in Figure 1.

Figure 1 Example Faces. **C**MU, **Y**ale, **N**ottingham

There are many other approaches to the face detection problem. Some techniques rely on a single face template or model for detection [10,5,8], others rely on facial subfeatures [12,13,11]. A variety of detection techniques are employed, from correlation [2], neural nets [8], creseptrons [14], eigentemplates [9,10,12], Baysian models [12], and flexible models [5,13]. Some approaches use bottom-up search [14], some use top-down search [1], and some combine both search types [9]. Combining bottom-up and top-down processes appears as a promising way to guide the search efficiently. There are systems in other domains which appear to use this approach successfully [6,7].

Some single face techniques [10] are advertised as generalizable to larger domains. Of the compound techniques using facial subfeatures [1], or those combining bottom-up and top-down search [6,9], the emphasis appears mostly on detection performance. I was interested to see what efficiency merits a compound and combined approach would have compared to a single template approach. To this end, I developed the Facets face detection system as a platform to compare both approaches in equivalent implementations.

In the compound template approach, a face is composed from a spatial model and four subfeatures: left eye, right eye, nose, and mouth. This two-level structure supports the combination of bottom-up and top-down search. Each subfeature type has an associated image analysis procedure used to detect the presence of the subfeature within the target image.

The search begins by invoking the image detection procedures for some set of subfeatures (e.g., left eyes). The choice of initial subfeatures critically affects accuracy and efficiency, and this is discussed in more detail below. When a subfeature is found, a face instance is created with the detected subfeature given as evidence. If the detected subfeature fits within the hypothesis space of an existing face instance, it is given to the existing instance. Managing overlapping hypotheses is discussed in more detail below.

A face instance represents a hypothesis about a face (or face-space) evidenced by the image, and it contains the subfeature instances already found (which provide evidence for the hypothesis). Parameters representing rotation, scale, and hypothesis strength are calculated based upon the evidence (subfeature instances) and default values. From these parameters the estimated feature space for missing subfeatures is determined.

The top-down search phase for a face instance begins after the newly created face instance is determined to represent a novel hypothesis. The image detection procedures are

called for the missing subfeatures using the estimated feature space as the search parameter. The search is scheduled using a priority queue. Searching and scheduling optimization issues are discussed below.

A face instance completes its search either when it has exhausted the subfeature search space, or when it represents a complete face instance. The entire search completes when the initial bottom-up search completes, and all face instances have completed their search.

TECHNIQUES

The foundation of the Facets face detection system is the simple template. A simple template is composed from a pyramid of masked grey scale images. Detection is performed using normalized correlation on the source image pyramid [3].

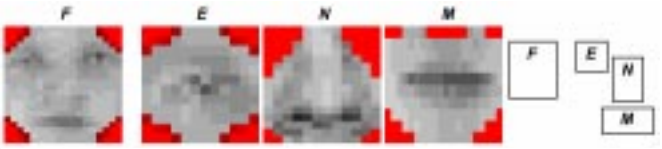

Figure 2 Feature templates with masks (left) and relative size (right). **F**ace, **E**ye, **N**ose, and **M**outh.

The target image pyramid is scaled by 75% for each level from bottom to top. For template pyramids, the scaling ensures that the maximum change in width or height for the scaled template is 2 pixels smaller than the preceding template in the pyramid. This ensures that templates are discretized to within a pixel of their neighbours, as required for correlation.

Feature and subfeature types are represented as a collection of template pyramids. The feature space for a template is composed from: the minimum and maximum area range in base coordinates, the rotation angle range, and the enclosing rectangle for the template center in base coordinates. Base coordinates are in pixels based upon the original (i.e. largest) target image.

Feature detection is performed by searching for a template over a target image within a given feature space. The initial candidate search traverses the image pyramid from top to bottom searching for the feature at the lowest resolution. At each level this is done by choosing templates from the template pyramid that are within the intersection of the given feature space and the feature space covered by the template pyramid, but not within the feature space previously searched. Using masked correlation, each template is compared against the current level's image, over the area specified in the feature space. If the correlation result is greater than the 0.7 threshold, the candidate feature space is recorded. From this localized candidate featurespace, the refine search continues down the pyramid looking for the highest resolution match. If the correlation score is over the threshold for the high resolution match, and if the resolution is sufficient, a feature instance is created.

The compound face uses a simple face model represented on a 2D plane with 5 degrees of freedom: the three rotational angles, the scaling factor, and the face center point. Faces contain subfeature collections for all subfeature instances within the hypothesis space for the face. Subfeature instances are added to the collection when a face is created, when overlapping faces are merged, and when detected features from the top-down search are returned to the requesting face instance. Redundant subfeatures are removed from the collection. When non-redundant subfeatures are added to the collection, a search is performed over the new subfeature combinations. The primary subfeatures are the combination of subfeatures which maximize the hypothesis strength. The estimated feature space is calculated for each subfeature type missing from the primary subfeatures.

A search request is scheduled if the feature space was not previously searched. The search priority is an estimate that the search will complete the face, and is based upon the hypothesis strength. The hypothesis strength is calculated as a weighted combination of the subfeature correlations scores, and subfeature distances from the ideal model in the estimated pose, to the actual subfeature locations.

The efficiency and accuracy of a combined search depends upon the subfeature space used for the initial bottom-up search. If the initial search comprises the entire subfeature space, then the search becomes purely bottom-up. If the initial feature space is not sufficient to find at least one subfeature per face, then the search cannot find all faces in the image. Our tests used the left eye subfeatures for the initial search.

When subfeature instances are detected, they are evidence for some face level hypothesis space. It is likely that overlapping hypotheses are created, especially if the initial subfeature search space is complete enough to ensure the detection of all faces. For efficiency it is important to merge similar face instances before they initiate their top-down search. If an existing face instance sufficiently overlaps the hypothesis space of the new face, the subfeatures of the new face are added to the existing face instance, and the new face instance is removed.

The scheduling queue provides a way to increase the feature space of the initial search without delaying the top-down phases. Searches are prioritized based upon face hypothesis strength and the phase of the aggregate search.

Aggregate search optimization combines the searches for several missing subfeatures and allows for control logic to guide and schedule searches. For example, cancelling or delaying the search for one subfeature if another subfeature

was not detected, or skipping the detection for remaining subfeatures if enough subfeature instances of that type were already found.

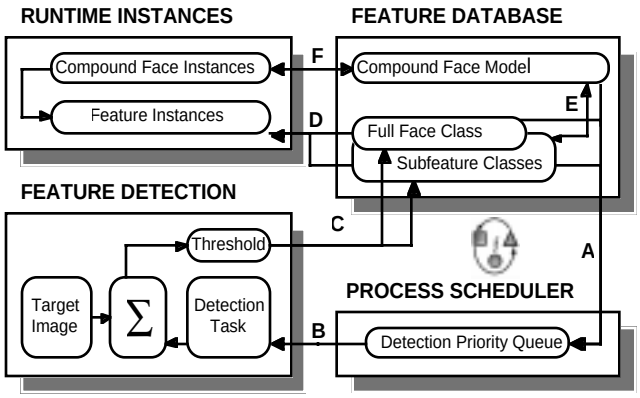

Figure 3 Facets Runtime Architecture.

During the runtime, as shown in Figure 3, requests for searches are sent (A) by the feature classes to the scheduler, where they are prioritized by value. Detection tasks are sent (B) to the detection unit to search for matches in the specified regions. Matches exceeding the threshold are returned (C) to the original template class. A subfeature instance is created and added (D) to the runtime instances. If the original search request came from a compound face, the instance is returned (E) to the face, otherwise a new compound face instance is created. Compound face instances are placed (F) among the runtime instances; but, to prevent searching overhead from multiple similar hypotheses, similar faces are merged. Compound face instances make localized search requests (F) via the compound face class, which communicates (E) with the subfeatures classes and then makes the request (A) using an aggregate search package.

DATA AND ANALYSIS

The template databases for the simple and compound technique were collected from the CMU 1104 image set. The training set includes 16 images with 25 faces each from 5 people, comprising 400 faces from 40 people in 10 poses all together. The database was created by collecting templates step by step so that newly acquired templates were not previously detected features. The simple template technique required 54 templates, totalling 2,010 KB in size, to cover the training set with 100% accuracy. By comparison, the compound template technique required 22 left eye templates (the initial detection templates), 44 right eyes, 25 noses, and 74 mouths, for a total of 165 templates, totalling 2,604 KB in size.

As more templates are added to the database, the domain coverage of the database increases. Tests were performed to determine the change in detection coverage and times as the database increases in size. The compound approach appears to fair better in both. Covering 60%, 70%, 80%, and 90% of the training domain requires a 9%, 14%, 18%, and 36% completeness of the compound subfeatures database, respectively, compared to a 19%, 26%, 33%, and 52% completeness of the simple template database. Both approaches are similar after 95% domain coverage. At 100% coverage, average detection times are 1.3 times slower for compound templates, but as new templates are added to the database the growth of detection times is linear for simple templates, while the growth is sublinear for compound templates.

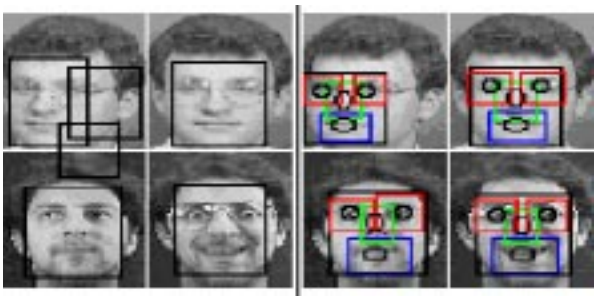

Figure 4 Detection examples for simple templates (left half) and compound templates (right half).

The performance of both techniques was evaluated on the Nottingham and Yale image sets using the template database acquired previously from the training set. On the Nottingham database the compound technique detected 93% of faces, compared to 70% for simple templates. For the Yale database the compound technique detected 35% of the faces, compared to 16% for simple templates. Compound templates also performed better in detection accuracy up to an order of magnitude.

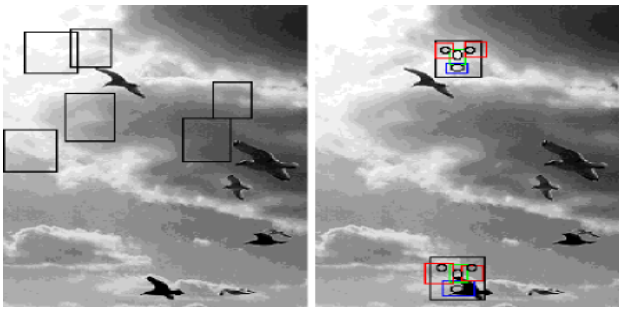

Figure 5 False positive example with simple templates (left) and compound templates (right).

To test false positive performance, both techniques were tried on three non-face images using the complete template databases. Compound templates made a total of 31 false positives, compared to 129 false positives for simple templates. Because the simple template approach is more likely to find smaller 'faces' than the compound approach, an adjustment was made to ignore faces below a thresholded size. With this adjustment the simple templates found 44 false positives.

Performance of both techniques was compared over images with varying scale and facial density. The test image was a 40 face subset from the training set. For the scaling tests, 9 scaled images were produced, ranging from 50% to 200% the size of the original image. For the facial density tests, 5 images were created by covering some of the faces with a textured swatch, producing images with 40 , 30, 20, 10, and 2 faces visible. Detection performance for compound templates decreased more abruptly away from the 100% image than for simple templates. The density tests demonstrated that compound templates are dependant upon feature density, since compound template detection times decreased linearly with reduced density, while simple template detection times remained constant.

Comparative tests were performed to determine the effectiveness of combining the bottom-up and top-down search, and of the aggregate search optimization. The search times for compound templates were about 30% slower than simple templates. Search times for all 165 subfeature templates were 4.5 times longer than the simple technique, and 3.5 times longer than the normal compound templates. Using aggregate search optimization sped up the compound technique by more than twice the time required for a complete top-down search.

CONCLUSIONS

Effective usage of available knowledge, both in the domain and about the image, appears a plausible way to make the detection process more efficient. Combining bottom-up and top-down search using subfeatures and models is one way to use partial image knowledge (features detected during bottom-up search) and domain knowledge (the modeled spatial relations localizing the top-down search). The face domain was chosen as a tractable, extensible, and interesting domain for study and the proposed technique was implemented in the Facets system.

Comparison tests were performed against an equivalent implementation of the simple template technique. The results appear to show that the compound template approach has merits worth further exploration. Except for the scale tests, compound templates performed better for both detection rates and accuracy. This may follow from the slightly higher resolutions used for the subfeature templates compared to face templates. As for the slower growth of the detection time as database size increases for compound templates, this may partially result from the merging of duplicate hypotheses. There is clearly a benefit from combining bottom-up and top-down search over bottom-up search alone, and an even further benefit from using aggregate optimizations that the top-down search affords. Overall, a 250% improvement in detection time. When compared to the top-down simple template technique, the compound approach lags slightly behind for feature dense images. Compound templates appear well suited though to images which are subfeature sparse, producing equivalent to slightly better detection times.

REFERENCES

[1] **Beymer**, David J. *Face Recognition Under Varying Pose* from *MIT AI Memo*, no 1461, 1993

[2] **Brunelli**, Roberto. Poggio, Tomaso. *Face Recognition: Features versus Templates* in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 15, no 10, pg. 1042-1052, 1993

[3] **Burt**, Peter J. *Smart Sensing within a Pyramid Vision Machine* in *Proceedings of the IEEE*, vol 76, no 8, pg. 1006-1015, August 1988

[4] **Jeng**, Shi-Hong. Liao, Hong. Han, Chin. Chern, Ming. Liu, Yao. *Facial Feature Detection using Geometrical Face Model: An Efficient Approach* in *Pattern Recognition*, vol 31, no 3, pg. 273-282, 1998

[5] **Lanitis**, Andreas. Taylor, Chris J. Cootes, Timothy F. *Automatic Interpretation and Coding of Face Images using Flexible Models* in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 19, no 7, pg. 743-755, 1997

[6] **Matsuyama**, Takashi. Hwang, Vincent Shang-Shouq. *SIGMA: A Knowledge-Based Aerial Image Understanding System*. Plenum Publishing Corporation, 199?

[7] **Milanese**, Ruggero. Gil, Sylvia. Bost, Jean-Marc. Wechsler, Harry. Pun, Theirry. *Integration of Bottom-Up and Top-Down Cues for Visual Attention Using Non-Linear Relaxation* in *Berkeley Technical Reports*, vol 94, no 14, pg. 1-6, 1994

[8] **Rowley**, Henry A. Baluja, Shumeet. Kanade, Takeo. *Neural Network-Based Face Detection* in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 20, no 1, pg.23-38, 1998

[9] **Shakunaga**, Takeshi. *Integration of Eigentemplate and Structure Matching for Automatic Facial Feature Detection* in *IEEE Journal - Unknown (0-8186-8344-9/98)*, pg. 94-99, 1998

[10] **Sung**, Kah-Kay. Poggio, Tomaso. *Example-Based Learning for View-Based Human Face Detection* from *MIT Lab Report*, pg. 1-8

[11] **Takacs**, Barnabas. Wechsler, Harry. *Detection of Faces and Facial Landmarks using Iconic Filter Banks* in *Pattern Recognition*, vol 30, no 10, pg. 1623-1636, 1997

[12] **Viola**, Paul A. *Complex Feature Recognition: A Bayesian Approach for Learning to Recognize Objects* from *MIT AI Memo*, no 1591. 1996

[13] **Yuille**, Alan L. Hallinan, Peter W. Cohen, David S. *Feature Extraction from Faces Using Deformable Templates* in *International Journal of Computer Vision*, vol 8, no 2, pg. 99-111, 1992

[14] **Weng**, John J. Hwang, Wey-Shiuan. *Toward Automation of Learning: The State Self-Organization Problem for a Face Recognizer* in *IEEE Journal - Unknown (0-8186-8344-9/98)*, pg. 384-389, 1998

Further Information

http://www.cs.ubc.ca/spider/gholst/MastersThesis