Outils stochastiques pour l'indexation vidéo

Stéphane Marchand-Maillet et Bernard Mérialdo
Département Communications Multimédia
Institut Eurécom - Sophia-Antipolis - France
{marchand,merialdo}@eurecom.fr

Résumé - Cet article évalue les capacités des Modèles de Markov Cachés (HMM) pour réaliser des tâches d'indexation vidéo. Nous nous intéressons plus particulièrement aux problèmes liés à la détection et localisation de personnes dans les séquences vidéo. Nous rappelons d'abord brièvement l'extension des modèles mono-dimensionnels au cas bi-dimensionnel et apportons les outils théoriques nécessaires au traitement de données continues. Nous détaillons ensuite les résultats de nos expérimentations en matière de localisation et suivi de visages à travers des séquences vidéo. L'utilisation de modèles stochastiques pour la localisation de visages permet une segmentation de l'image et donne aussi une mesure (que l'on peut qualifier de continue) quant à la qualité de cette segmentation. Dans une troisième partie, nous utilisons cette mesure pour réaliser la détection de personnes dans la séquence vidéo. Nous montrons finalement que cette mesure de détection peut être exploitée pour réaliser une segmentation de la vidéo. Dans ce cas, nous montrons qu'une forme de micro-segmentation peut-être obtenue à partir de notre mesure. L'article est conclu par une discussion sur les potentialités des outils stochastiques pour l'indexation vidéo.

1 Introduction

La modélisation stochastique s'avère être un outil puissant pour l'automatisation des techniques de reconnaissance de signaux. Par exemple, les Modèles de Markov Cachés mono-dimensionnels (1D-HMM) sont couramment utilisés avec succès dans les problèmes de reconnaissance de la parole [7, 8] . Ils se montrent performants par leur flexibilité et efficacité, tant en termes de modélisation qu'en termes de calcul. Toutefois, peu de travaux ont effectivement appliqué ces modèles dans le contexte de reconnaissance de signaux 2D. Quelques exemples étudient la reconnaissance de visages [9] et la localisation de mots-clés dans des images de documents [2, 3].

Dans cet article, nous nous proposons d'examiner plus avant l'utilisation des modèles de Markov Cachés pour l'analyse de séquences vidéo. La section 2 présente brièvement la modélisation stochastique de données bi-dimensionnelles. L'idée d'utiliser une structure *pseudo*-2D est basée sur le fait que la connectivité 2D est inutilisable en pratique car elle mène à une complexité exponentielle de la masse de calcul à effectuer [4]. Les Modèles de Markov Cachés pour le cas Pseudo-2D (P2DHMM)

ont été présentés à l'origine pour des images binaires [3]. Dans ce travail, nous étendons les P2DHMM pour traiter une information de couleur. Nous présentons alors trois applications à l'analyse de séquences vidéo où les P2DHMM offrent des propriétés intéressantes (section 3). Tout d'abord, nous exploitons la segmentation des images définie implicitement par la séquence d'états la plus probable associée à une séquence d'observations. Dans ce cas, le P2DHMM constitue un modèle déformable facile à manipuler. La localisation de visages est présentée comme étant un cas particulier de cette segmentation. Finalement, en utilisant la probabilité de chaque modèle comme une mesure, nous montrons comment une information utile peut être extraite dans le but de l'indexation vidéo et la sélection automatique d'image-clés [1].

La section 4 conclu l'article par une discussion sur les avantages et inconvénients induits par l'utilisation des P2DHMM pour ce type d'applications.

2 Modèles de Markov Cachés pseudo 2D

Les HMM [7, 8] sont des modèles stochastiques qui offrent un haut niveau de flexibilité pour la modélisation de séquences d'observations. Ils permettent de recouvrer la structure (cachée) de cette séquence d'observations en associant à chaque observation un certain état (caché). Le temps passé dans chaque état n'étant pas contraint, ceci permet aux Modèles de Markov Cachés de réaliser une association entre un modèle donné et la séquence d'observations en question (elastic matching). La structure du modèle utilisé est alors simplement définie par les contraintes imposées sur les transitions entre états. Dans cette section, nous présentons une généralisation des 1D-HMM pour la modélisation de données bi-dimensionnelles.

2.1 Principe

La causalité dans un ensemble d'observations (pixels) $O = \{o_{xy}\}_{x=1...X,y=1...Y}$ bi-dimensionnel peut être définie comme étant la dépendance entre o_{xy} et son voisinage. Toutefois une structure complète pour la connectivité résulte en la définition de problèmes NP-complets pour l'évaluation des paramètres du modèle (entraînement) et pour retrouver la structure cachée associée à une séquence d'observations (segmentation) [4]. Dans cet article, nous choisissons d'utiliser uniquement la connectivité entre lignes. Une ligne $O_y = \{o_{1y}, \ldots, o_{Xy}\}$ est considérée comme étant elle-même une observation de haut niveau et la séquence $O = \{O_1, \ldots, O_Y\}$ de lignes est modélisée par un 1DHMM contenant des *super*-états. La figure 1 montre un exemple de P2DHMM Λ qui est une extension directe d'un modèle de Bakis classiquement utilisé en reconnaissance de parole.

La structure d'un P2DHMM modélisant une séquence bi-dimensionnelle d'observations sera composée de différents 1DHMM (horizontaux, un par type de ligne) connectés par un modèle vertical représentant les relations entre les lignes.

La séquence (cachée) Q d'états la plus probable associée à O connaissant le modèle Λ sera décrite à deux niveaux. Q est tout d'abord une séquence de super-états $\{Q_y,\ y = 1, \ldots Y\}$, chacun indiquant le 1DHMM correspondant à la ligne

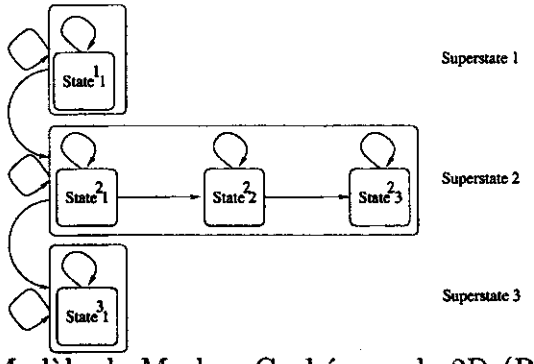

Figure 1: Modèle de Markov Caché pseudo-2D (P2DHMM).

d'observations O_y. Chaque super-état Q_y est composé d'états q_{xy} ($Q_y = \{q_{1y}, \ldots, q_{X_y}\}$) indiquant l'état du 1DHMM correspondant à la position (x, y) (*i.e.*, correspondant à l'observation o_{xy}).

2.2 Développements

Un P2DHMM peut être formalisé par la description $\Lambda = \{\lambda, A, B, \Pi\}$, où, $\lambda = \{\lambda^i \ ; \ i = 1, \cdots, N\}$ est l'ensemble des N super-états possibles dans le modèle. Chacun des ces super-états est un 1DHMM λ^i avec les paramètres $\lambda^i = \{s^i, V, A^i, \pi^i\}$:

- $s^i = \{s^i_1, \ldots, s^i_{N^i}\}$ est l'ensemble des N^i états possibles et $V = \{v_1, \ldots, v_L\}$ est le vocabulaire de sortie, commun à tous les super-états.
- $A^i = \{a^i_{kl} = P[q_{xy} = s^i_l | q_{x-1y} = s^i_k]\}_{k,l=1\ldots N^i}$ est l'ensemble des probabilités de transition à l'intérieur du super-état λ^i.
- $B^i = \{b^i_k(l) = P[o_{xy} = v_l | q_{xy} = s^i_j]\}_{\substack{k=1\ldots N^i, \\ l=1\ldots L}}$ est l'ensemble des probabilités d'émission du super-état λ^i.
- $\pi^i = \{\pi^i_j = P[q_{1y} = s^i_j | \lambda^i]\}_{i=1\ldots N}$ est l'ensemble des probabilités initiales de chaque état du super-état λ^i.

$A = \{a_{i,j} = P[q_y = \lambda^j | q_{y-1} = \lambda^i]\}_{i,j=1\ldots N}$ est l'ensemble des probabilités de transitions entre super-états.
$\Pi = \{\pi_1 = P[Q_1 = \lambda^i | \Lambda]\}_{i=1\ldots N}$ est l'ensemble des probabilités initiales pour chaque super-état.

Dans [3], l'entraînement est basé sur la modification de l'algorithme de Viterbi pour le cas pseudo-2D. Dans notre étude, nous avons étendu la procédure de Baum-Welsh pour la réévaluation des paramètres de notre P2DHMM dans le cas de données continues (voir [5] pour le détail). Ceci est motivé par le fait que bien que discrètes les observations vont parcourir un espace trop grand pour être géré par des histogrammes ($L = 256^3$, par exemple). Dans ce cas, les probabilités d'émission sont modélisées par des multi-Gaussiennes de la forme:

$$b^i_j(o_{xy}) = \sum_{m=1}^{M^i_j} \frac{c^i_{jm} \exp\left(-\frac{1}{2}(o_{xy} - \mu^i_{jm})^T (\Sigma^i_{jm})^{-1} (o_{xy} - \mu^i_{jm})\right)}{\sqrt{(2\pi)^D |\Sigma^i_{jm}|}},$$

où D est la dimension de chaque observation, M^i_j le nombre de mixtures $\mathcal{N}(o_{xy}, \mu^i_{jm}, \Sigma^i_{jm})$ et c^i_{jm} les coefficients de mixtures pour l'état s^i_j. La réestimation des paramètres de

ces distributions est aussi comprise dans notre procédure.

2.3 Outils de modélisation

En résumé, nos développements nous fournissent trois opérations de base que nous allons combiner pour différentes applications.

• Entraînement. Étant donnée une topologie de transition entre états dans un modèle Λ et l'ensemble des images $\{I_1, \ldots, I_m\}$, en itérant la réestimation des paramètres du modèle Λ grâce à la procédure de Baum-Welsh, nous définissons les paramètres qui maximisent $\Pi_{i=1}^{m} P[I_i|\Lambda]$. L'initialisation de cet entraînement pouvant être donné ou aléatoire.

• Mesure de performance du modèle. Soit un modèle donné Λ, décrit par ses paramètres, on peut estimer $P[I|\Lambda]$, la mesure de correspondance du modèle par rapport à une séquence d'observation I grâce à la procédure de Viterbi ou comme résultat de la procédure de Baum-Welsh.

• Segmentation contrainte. Soit un modèle donné Λ, décrit par ses paramètres, en utilisant une procédure de Viterbi (comme celle décrite dans [3]), on peut trouver Q, la séquence d'états la plus probable par rapport à l'image I. Cette séquence d'états défini implicitement une segmentation de l'image I contrainte par la topologie du modèle Λ.

3 Applications à l'analyse de séquence vidéo

Nous illustrons maintenant l'utilisation de ces trois opérations dans le cadre de l'analyse de séquence vidéo. Trois applications complémentaires sont étudiées dans le contexte des P2DHMM.

3.1 Suivi de région (warping)

Dans cet exemple, la segmentation contrainte est utilisée pour suivre les déformations d'une image à travers une séquence vidéo. L'image est segmentée grâce à un P2DHMM contenant 7 super-états de 7 états chacun. L'image sera donc segmentée en une grille 7×7 de 49 états. Le modèle est entraîné grâce à des exemples du type d'image à étudier. L'initialisation est simplement faite en partitionnant l'image en 49 régions rectangulaires régulières. A la convergence, le modèle aura choisi la configuration d'états la plus probable par rapport à toutes les images d'entraînement.

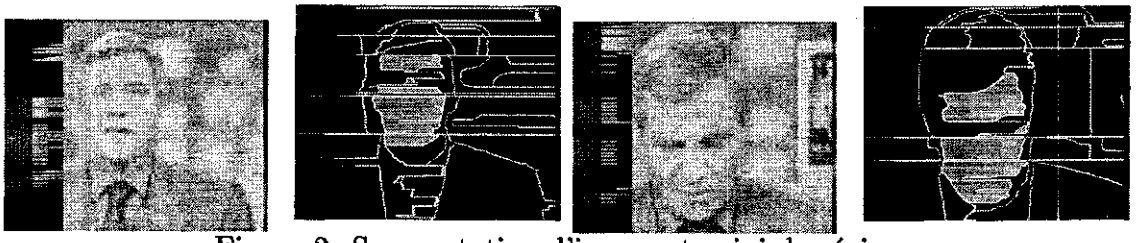

Figure 2: Segmentation d'image et suivi de région.

Les contours des segmentations obtenues sont illustrés en figure 2. Les régions correspondant aux états 16, 30 et 44 (de gauche à droite et de haut en bas) sont colorées pour les deux images. Ceci montre une forme de suivi de régions dans l'image et donc illustre la capacité des P2DHMM à être utilisés comme modèles déformables pour le suivi de région.

3.2 Localisation de visages

Basé sur le principe général de segmentation contrainte par un modèle P2DHMM, nous montrons maintenant comment un modèle simple peut être construit pour réaliser la localisation d'un visage dans une image. Le P2DHMM (Λ_{Vis}) décrit en figure 1 est utilisé pour modéliser l'image contenant un visage schématisée ci-dessous (figure 3). Pour une étude plus détaillée, le lecteur est renvoyé à [6].

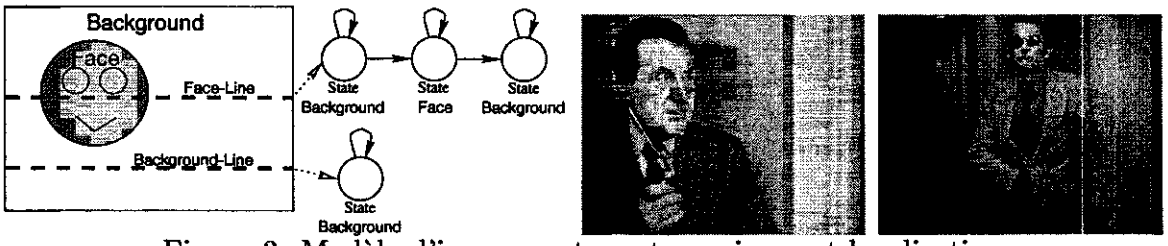

Figure 3: Modèle d'image contenant un visage et localisation.

Pour assurer la cohérence dans le modèle, les probabilités d'émission de tous les états **Background** doivent être égales. Ceci est réalisé grâce à la définition de deux distributions **Foreground** et **Background** associées à leurs états respectifs via la notion de pointeur.

L'entraînement est fait grâce à la procédure de Baum-Welsh basée sur un ensemble d'images d'entraînement segmentées à la main. La segmentation finale est donnée par la procédure de Viterbi. La figure 3 donne des exemples de résultats obtenus à partir d'images d'une séquence vidéo de type "Fiction", les pixels correspondant au visage (état **Foreground**) ont été blanchis dans ces images.

Dans les séquences vidéo, les contraintes classiques d'orientation, de taille et de pose ne peuvent pas être appliquées. Un avantage donné par les P2DHMM pour la localisation de visages est de permettre la localisation en modélisant le fond de l'image statistiquement et en ne se basant pas sur les caractéristiques du visage mais plutôt sur les statistiques globales de la région.

Un autre avantage des P2DHMM est le fait qu'ils donnent une segmentation précise de l'image en fin de processus. Dans notre cas, les visages peuvent donc être extraits de l'image originale, ce qui va faciliter les opérations de reconnaissance suivant la localisation (figure 4).

3.3 Analyse de séquences vidéo

Les applications ci-dessus utilisent principalement la segmentation donnée par la procédure de reconnaissance. Dans cette section, nous utilisons la performance d'un modèle ($P[O|\Lambda]$) comme une mesure (continue) pour évaluer les similarités entre images successives.

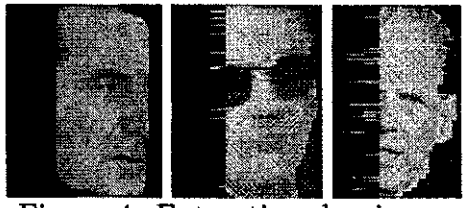

Figure 4: Extraction du visage.

Détection d'évènement La segmentation d'une séquence vidéo est définie comme le partitionnement de cette séquence en plans. Soit un modèle Λ pour un type particulier d'images, la valeur de $P[I|\Lambda]$ pour toutes les images de la séquence permet de segmenter la séquence en évènements (*micro*-segmentation). Nous présentons ce type d'applications à travers l'exemple de localisation de visages ci-dessus.

Le modèle Λ_{Vis} est utilisé et la valeur de $P[I|\Lambda_{\text{Vis}}]$ est comparée à celle correspondant à un modèle d'une image générique $P[I|\Lambda_{\text{Fond}}]$. Suivant notre approche, le modèle Λ_{Fond} est tout simplement un modèle contenant 1 super-état d'un état, ce qui n'est rien d'autre qu'un simple histogramme créé à partir d'exemples d'images ne contenant pas de visage. Cette mesure est calculée pour toutes les images à travers la séquence et le graphe est montré en figure 5. Les images correspondant aux point-clés de ce graphe sont aussi montrées pour illustrer la correspondance.

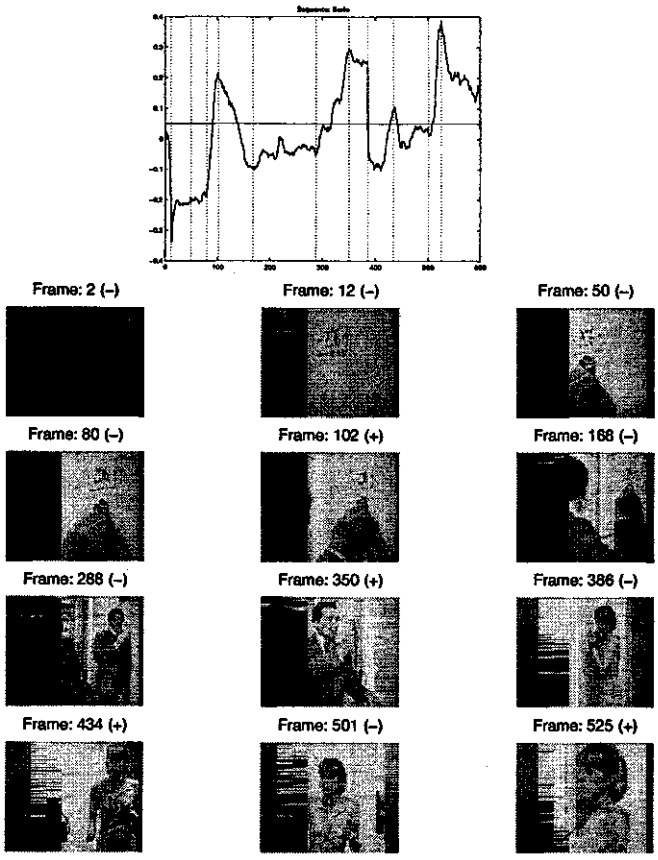

Figure 5: Détection de visages.

Ce graphe montre comment notre mesure peut être exploitée pour caractériser une information à partir de la séquence en question. Par exemple, entre les images 80

et 168, une personne apparaît et tourne la tête. Du fait de l'utilisation du modèle Λ_{Vis}, on observe un maximum de notre mesure durant cette sous-séquence. Les images 80 (début), 102 (maximum) et 168 (fin) peuvent donc être choisies comme image-clés pour décrire le contenu de la séquence.

La détection de visages peut être obtenue en seuillant ce graphe par une valeur définie durant l'entraînement. Schématiquement, des exemples d'images contenant ou non un visage sont présentées au modèle et la valeur τ qui sépare les mesures correspondant à ces deux classes d'images est utilisée comme seuil. Dans notre cas, nous obtenons comme valeur $\tau = 0.04$, ce qui conduit aux détections illustrées par des signes + et - au-dessus de chaque image.

De plus, comme nous utilisons le même modèle pour toutes les images de la séquence, notre mesure permet la comparaison d'images consécutives dans la séquence. Le gradient de cette mesure devrait donc mettre en avant les changements dans la séquence et montrer les coupures (cut detection). Ceci est démontré en figure 6 qui compare les différences entre images consécutives en utilisant les valeurs du graphe ci-dessus (figure 6-gauche) et les valeurs de différence d'histogrammes (figure 6-droite).

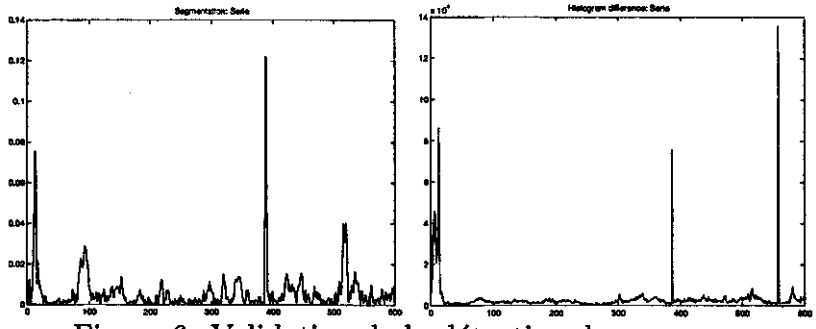

Figure 6: Validation de la détection de coupures.

Sélection d'image-clés La généralisation du processus de micro segmentation pour la sélection d'image-clés peut être définie comme suit. Soit une séquence de m images consécutives $\{I_i\}_{i=1...m}$, un modèle Λ est entraîné grâce à ces images et l'image I^* qui correspond le mieux (en termes de probabilités) à ce modèle est choisie comme image-clé de la séquence. Cette technique assure que la séquence sera représentée par une image correspondant à l'image la plus fréquemment rencontrée durant la séquence.

4 Discussion et conclusion

Dans cet article, nous avons introduit l'utilisation des modèles stochastiques pour l'analyse de séquences vidéo à travers la présentation des P2DHMM. Nous avons aussi présenté trois applications où les P2DHMM se révèlent performants comme compléments aux techniques utilisées classiquement.

L'avantage de tels modèles se situe dans la simplicité et la flexibilité de la modélisation. Bien que la vraie structure bi-dimensionnelle des données ne soit pas com-

plètement prise en compte, notre expérience montre que ces modèles sont suffisants pour réaliser certaines tâches d'indexation. De plus, utilisant des techniques à base d'entraînement, ces modèles sont associés à des procédures qui permettent le calcul de paramètres optimaux par rapport à une base d'entraînement donnée. Ces outils forment donc un ensemble d'outils cohérent pour le développement d'applications dans le contexte de l'indexation et l'analyse de données vidéo.

Une des faiblesses majeures de ces outils est leur complexité algorithmique qui rend le calcul en temps réel impossible s'ils sont utilisés tels que. De plus, pour certains exemples d'applications, des moyens plus simples se montrent déjà très performants. Ces outils ne doivent donc pas être utilisés aveuglément en remplacement d'outils existants mais plutôt choisis et adaptés dans un contexte où le manque de contraintes ne permet pas d'exploiter d'autres options plus classiques.

De par notre expérience, nous nous attachons à développer des outils de modélisation stochastique pour l'indexation vidéo. Nous voyons ces outils comme une base de travail dans laquelle il est possible de définir des outils performants qui aideront à résoudre des problèmes mal traités actuellement.

References

[1] F. Idris and S. Panchanathan. Review of image and video indexing techniques. *Journal of Visual Communication and Image Representation*, 8:146–166, 1997.

[2] S.-S. Kuo and O. E. Agazzi. Automatic keyword recognition using Hidden Markov Models. *Journal of Visual Communication and Image Representation*, 5(3):265–272, 1994.

[3] S.-S. Kuo and O. E. Agazzi. Keyword spotting in poorly printed documents using Pseudo 2-D Hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-16(8):842–848, 1994.

[4] E. Levin and R Pieraccini. Dynamic planar warping for optical character recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume III, pages 149–152, 1992.

[5] S. Marchand-Maillet. 1D and pseudo-2D Hidden Markov Models for image analysis – A: Theoretical introduction. Technical Report RR-99-49, Institut EURECOM, Dept of Multimedia Communications, 1999.

[6] S. Marchand-Maillet and B. Mérialdo. Pseudo two-dimensional Hidden Markov Models for face detection in colour images. In *Proceedings of the Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, Washington DC, USA, 1999.

[7] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

[8] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceeding of the Second IEEE Workshop on Applications of Computer Vision*, Sarasota, Florida, December 1994.