BAYESIAN CONCEPTS FOR HUMAN TRACKING AND BEHAVIOR DISCOVERY

Ing. Petr CHMELAŘ, Doctoral Degree Programme (1)
Dept. of Information Systems, FIT, BUT
E-mail: chmelarp@fit.vutbr.cz

Supervised by: Dr. Jaroslav Zendulka

ABSTRACT

This work describes a framework for vision based human detection, tracking, pose recognition and behavior discovery in a uniform manner of Bayesian classifiers. It considers mathematical concept of identification by classification of preprocessed images and regression of poses for discovering behavior patterns in its databases using Bayes' nets.

1 INTRODUCTION

The main reason why to use probabilistic modeling for identification is the need for unified mathematic formulation in different areas. We concern classification of hidden parameters, which is affected by information loss, illumination noise and segmentation errors what can be understood as a noise in a communication channel in figure 1.

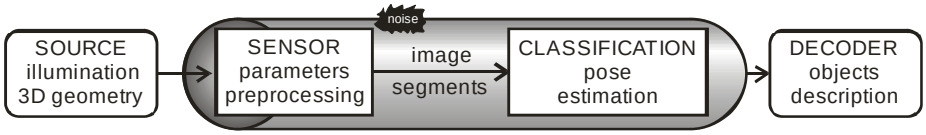

Fig. 1: *Scene description as a channel (from [2])*

The **visual information** and included **prior knowledge** of the sensed world are the main resources of considered application. The image acquisition and preprocessing is followed by the object recognition and its' pose estimation. Also these techniques may misclassify sensed objects due to incompleteness or occlusions. Poses and tracks are stored in database. Mining these spatio-temporal data provides useful information about sensed world and can influence other modules of the vision system by discovering frequent patterns of behavior used than as a posterior to prior knowledge feedback.

2 COMMON BASE

The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the chance of the thing expected upon sic (it's) happening. (from [6])

The Bayes' definition of probability looks limited nowadays because assumes only the extent of observable consequences. But it is well suitable for us. Consider following system:

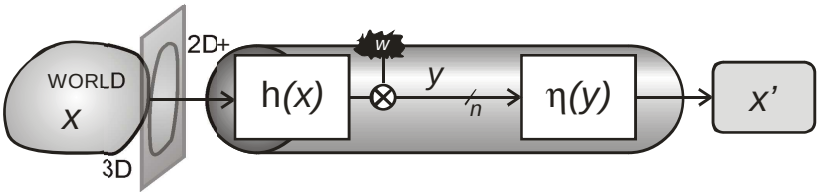

Fig. 2: *Parameter estimation system*

In figure 2 parameter $x \in X = \{x_1, x_2, \dots x_k\}$ is a **hidden state** or class that cannot be observed directly, $y \subseteq Y = \{y_1, y_2, \dots y_n\}$ is an **observation**, information we can get about real world. In computer vision it is usually a result of preprocessing and segmentation of sensor data, represented by function $h(x)$, the encoder. For instance if x is an *apple*, $y = h(x)$ could be *(round, red).* We cannot be sure because the channel is encumbered by a **white noise** w. In very general it holds *complete information = observable information + information loss*.

The **identification problem** requires a decoder – mapping $\eta(y)$ between Y and X' that is often called **classification** or **parameter estimation** function. We presume the result of identification $x' \in X' = \{x'_1, x'_2, \dots x'_k\}$ is an **optimal estimation** of class c or parameter θ corresponding to x, informally $x \sim c \sim \theta_c$. Due to our goal, developing reliable identification system we seek for classifiers with minimum **estimation error** rates.

That's the time to introduce **Bayes' theorem** which is derived from conditional probability $P(X \mid Y)$, the probability of event X given event Y is

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} \quad and \quad P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}, \tag{1}$$

where $P(Y \mid X)$ is the likelihood of Y given X. We can find that

$$P(X \mid Y)P(Y) = P(X \cap Y) = P(Y \mid X)P(X), \tag{2}$$

presuming that probability $P(Y) > 0$, we obtain Bayes' theorem:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} \tag{3}$$

Each term has a conventional name. $P(X)$ is the **prior** information having no information about $P(Y)$, which is the prior **marginal** probability, acting as a normalizing constant and can be counted as the sum of all mutually exclusive hypotheses $\sum_X P(Y \mid x_i)P(x_i)$. $P(Y \mid X)$ is **likelihood** or posterior probability given by the system or training. Finally $P(X \mid Y)$ is the **posterior** probability, the conditional probability of X is derived from Y. Within this terminology the theorem can be rephrased as the normalized likelihood multiplied by prior probability and it provides a method for adjusting degrees of belief of new information.

3 CLASSIFICATION

In the case of classification we define a **loss function** $\lambda(x_1, x_2)$ that penalizes classification errors of observation belonging to class x_1 to x_2 of X. We take for granted that correct decisions are cheaper than misclassifications. We choose the classification rule $\eta^*(y)$ that optimizes expected classification loss by minimization

$$\eta^*(y) = \arg\min_{\eta(y)} \sum_X \lambda(x_i, \eta(y)) P(x_i \mid y) \tag{4}$$

where $P(x \mid y)$ is the posterior probability for observing class x given y. Having especially *0-1* loss function charging classification errors by *1*, the loss function fade out the highest summand. Therefore we determine the highest posterior probability as

$$\eta^*(y) = \arg\max_x P(x \mid y) = \arg\max_x P(y \mid x) P(x) \tag{5}$$

and the optimal decision rule is called **Bayesian classifier.**

3.1 NAIVE BAYES CLASSIFICATION

There are two possibilities how to make inferences about parameters of the underlying probability distribution of a given data set in (5). **Maximum likelihood estimation** or MLE simply expressed by maximal $P(x \mid y)$ and its regularization known as **Maximum a posteriori** or MAP, presented as maximum of $P(y \mid x)P(x)$.

If we presume a training dataset D with categorical data, class prior probabilities of class $x \in X$ is $P(x) = |D_x| / |D|$ that means relative count of samples of class x in its collection D as in [1]. Using this prior probability leads Bayesian to interesting data mining techniques, similar to association rules with they're support and confidence.

In spite of capturing more concise structure information about analyzed data, the direct training of **joint probabilities** of all observations Y prone extremely to **overfitting**. The solution is the **naive assumption** that each y is **conditionally independent**

$$y_1 \perp y_2 \quad \Leftrightarrow \quad P(y_1 \mid y_2) = P(y_1) \quad or \quad P(y_1 \wedge y_2) = P(y_1) P(y_2) \tag{6}$$

or that attributes are not correlated. Thus we can get MAP as maximal $P(y \mid x)P(x)$

$$P(y_1 \wedge y_2 \wedge \ldots y_n \mid x) = \prod_{i=1}^{n} P(y_i \mid x) \quad where \quad P(y_i \mid x) = \frac{|D_{x \wedge yi}|}{|D_x|} \tag{7}$$

Example 1: Presume that there can lie *apple, orange* or *chocolate* on the table, $P(x) = 0.33$. We know that *apple* is *(round, red)*, *orange* is *(round, orange)* and *chocolate* is *(rectangle, brown)*. Thus we have $y_1 = \{round, rectangle\}$ and $y_2 = \{red, orange, brown\}$. If we want to classify red rectangle we count the highest $P(y \mid x)P(x)$, where each $P(y \mid x)$ is computed from training data:

$P(rectangle \mid apple) * P(red \mid apple) * P(x) = 0.1 * 0.6 * 0.33 = 0.0198$

$P(rectangle \mid orange) * P(red \mid orange) * P(x) = 0.1 * 0.3 * 0.33 = 0.0099$

$P(rectangle \mid chocolate) * P(red \mid chocolate) * P(x) = 0.8 * 0.1 * 0.33 = 0.0264$

We can see that maximal posterior probability has the (unpacked) red chocolate.

4 REGRESION

Up to now we considered the mapping $\eta(y)$ as a function predicting discrete categorical output. Real valued representation called **regression** is more suitable in parameter estimation. For instance the pose of a rigid object consists of *3* translational and *3* rotational degrees of freedom in the world coordinate system, denoted as $\theta \in \boldsymbol{R}^6$. Regression function

$\eta(y) = \theta_x$ depends on the actual class *x*. The most commonly used loss function is the square error $\lambda(\theta_x, \eta(y)) = |\theta_x - \eta(y)|^2$. The regression function is generally the minimization

$$\eta^*(y) = \arg\min_{\eta(y)} \int \lambda(\theta_x, \eta(y)) p(\theta_x \mid y) d\theta \tag{38}$$

where $p(\theta_x \mid y)$ is the probability **model density** function of θ_x given *y*. Similar to Bayesian classifiers is the **conditional expectation** $\eta(y) = E[\theta_x \mid y]$. Other representations of regression in statistics are parametric (linear) functions that restrict the parametric family and regularization to avoid overfitting. For further information see [2].

There is a great deal to accomplish pose estimation in computer vision. We should incorporate all available knowledge into the model construction like its size, shape or 3D structure where the transformation of image to the world coordinate system is available as well as the prior probability density of objects' localizations and tracks. Also to the human recognition process should embed methods for accurate motion estimation covered by texture analysis for successful segmentation and recognition of basic shapes.

Assuming uniform parameter distribution, the optimal estimation is then maximization

$$\theta_x^* = \arg\max_{\theta x} p(\theta_x \mid y) \tag{49}$$

but we can presume more in our case. For example pedestrian location depends on sidewalks and passages outside or doors inside buildings. Orientation is given by they're track. Also a human body can be split into *10-14* parts (head, torso and twice upper arm, forearm with hand, thigh and calf with foot). Its' sizes are derived from biometric expectation based on the measured height, so the pose recognition falls to detect angles of *10* rectangles.

5 BAYESIAN NETS

The main disadvantage of methods described above is the assumption that classes or parameters are conditionally independent to prevent overfitting. In general we cannot make such presumption. **Bayesian belief networks** specify both categorical and continuous joint conditional probabilities in a natural way using graph theory.

It consists of a graph in which each node *Z* represents a random variable *z*, $z \sim y$ when *z* is **observable** or $z \sim x$ when is **hidden,** representing a missing value. Edges represent a probabilistic dependence of nodes. Bayesian nets are in general oriented graphs. If there is an edge from Z_1 to Z_2, the node Z_1 is called **predecessor** or parent of Z_2 called **descendant**. Each variable is independent of its ancestors given its parents, where the ancestor relationship is with respect to some fixed topological ordering of the nodes.

The second component defining a belief network is a **conditional probability distribution** or CPD. That is a list that for each node *Z* specifies the conditional distribution *P(Z | Parents(Z))*, probability of node *Z* given by combination of all its parents.

Example 2: There are *4* nodes with categorical variables in figure 3. We don't know whether it rained or the sprinkler watered the grass, which is evidently wet. Our decision can influence observation of the sky. If it is cloudy there is higher possibility that rained than if the sky is bright. It can be shown in a manner of Bayesian belief networks.

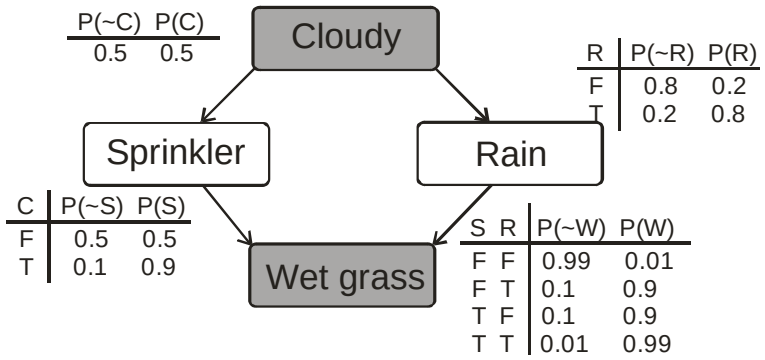

Fig. 3: *Example of Bayes net: We can see that the grass is wet. (from [5])*

5.1 TEMPORAL BAYES NETWORKS

Dynamic Bayesian networks or DBN are directed graphical models representing the hidden state in terms of state variables, which can have complex interdependencies. Simple kind is **Hidden Markov Models**, which has one discrete hidden node and one continuous observed node per time slice, *4* in fig. 4, showing dark circles as continuous observable nodes; squares denote discrete and white hidden states. Also **Kalman filter** is powerful [3] prediction-correction real-valued kind of DBN that can model even more than object tracking.

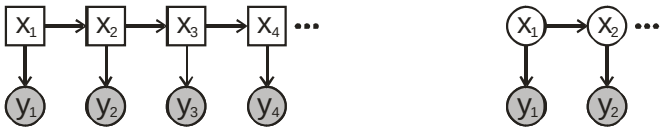

Fig. 4: *Illustration of Hidden Markov and Kalman filter models*

CONCLUSION

This work briefly introduces pattern recognition methods using Bayesian probability modeling. Bayesian classification in necessary in object or human recognition and in area of data mining for labeling unknown objects [1]. Regression is useful for pose estimation and tracking in real-world coordinate system [2]. It can use also advantages of Kalman filter, ensuring the smoothness of moving [3]. Tracks are saved to database and used as a learning sample for temporal Bayesian nets, that are well suitable [4] to detect nontrivial human behavior and its mutual influence either statistically dominant or outline. And that can be finally used as prior information to provide higher quality detection and recognition.

REFERENCES

[1] Han, J., Kamber, M. Data Mining: Concepts and Techniques. 2001. ISBN 1-55860-489-8.

[2] Jähne, B., Haussecker, H., Geissler, P.: Handbook of Computer Vision and Applications. 1999. ISBN 0-12-379770-5.

[3] Sorenson, H. W.: Least-squares estimation: from Gauss to Kalman. 1970.

[4] Olivier, N., Rosario, B., Pentland, A.: A Bayesian Computer Vision System for Modeling Human Interactions. 1999.

[5] Murphy, K.: A Brief Introduction to Graphical Models and Bayesian Networks. 1998.

[6] Wikipedia, the free encyklopedia. 2006.