# DIRICHLET-TREE DISTRIBUTION ENHANCED RANDOM FORESTS FOR FACIAL FEATURE DETECTION

*Yuanyuan Liu*[1,2,3] *, Jingying Chen*[*1,2], *Cunjie Shan*[1*]

[1]National Engineering Research Center for E-Learning, Central China Normal University,Wuhan,China
[2]Collaborative & Innovative Center for Educational Technology (CICET)
[3] Wenhua College, Wuhan, China

## ABSTRACT

A cascaded approach for facial feature detection in unconstrained environment, e.g., various poses and illuminations, occlusion, low image resolution, different facial expressions and make-up, is proposed in this paper. First the positive facial area is extracted to eliminate the influence of noise under various conditions. Then a Dirichlet-tree distribution enhanced random forests (D-RF) algorithm is proposed to detect facial features using cascaded head pose models in local sub-regions. Meanwhile, multiple probability models are learned and stored in leaves of the D-RF, i.e., a positive/negative patch probability, a head pose probability, the locations of facial features and facial deformation models (FDM). Finally, the composite weighted voting that fuses classification and regression methods is used to decide the locations of facial features. Experiments with the public databases demonstrate the robustness and accuracy of the proposed approach.

***Index Terms***— D-RF, Head pose estimation, Facial feature detection, FDM, Composite weighted voting

## 1. INTRODUCTION

Facial feature detection is often the first step for many applications such as face recognition, facial expression analysis and visual focus of attention recognition. Most of the existing methods [1, 2, 3, 4, 5] focus on the facial feature detection in constrained environment. However, facial feature detection still remains a challenge in unconstrained environments.

In recent year, Random Forest (RF) is a popular method in computer vision given their capability to handle large training datasets, high generalization power and speed, and easy implementation [6, 7, 3, 8]. Furthermore, Matthias. *et al.* proposed a conditional random forest (C-RF) to detect facial features under 5 horizontal head poses [3]. The accuracy rate reaches 72.15% in natural environment. In order to improve the accuracy and efficiency in unconstrained environment, we introduce a Dirichlet-tree distribution algorithm into RF framework. Minka proved the high accuracy and efficiency of the Dirichlet-tree distribution in [9]. Some researchers used a Dirichlet-tree distribution in multi-objects tracking and affective computing [6, 10]. In this work, the D-RF is proposed to detect facial features in unconstrained environment. The flowchart of the proposed D-RF is given in Fig.1, where D-f1 and D-f2 are the cascaded layers in the D-RF.

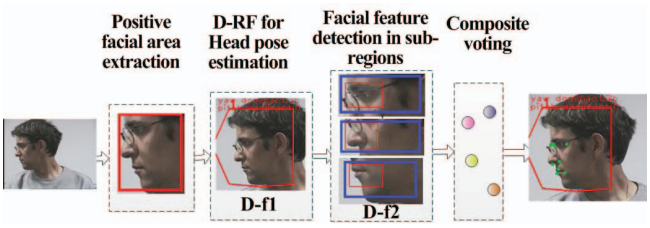



**Fig. 1**. Facial feature detection using the D-RF.

The main contributions of this paper are as follows. First, the positive facial area is extracted to eliminate the influence of noise under various conditions. Then local sub-regions are searched for using Adaboost [7] with Haar-like features to detect local features. Furthermore, multiple probability models are learned and stored in leaves of the D-RF. Finally, to deal with the imbalanced distribution of samples and multiple leaf models, the composite weighted voting method is used to decide the final locations of facial features. The experiment results show good performance of the proposed approach.

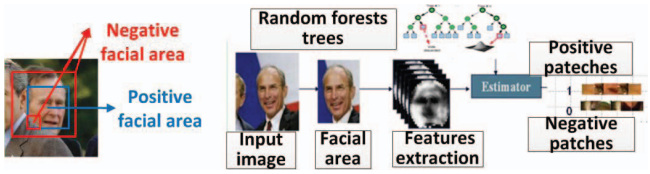
## 2. POSITIVE FACIAL AREA EXTRACTION

The extracted facial area using Jones &Viola detector [11] usually includes some noise, such as hair and occlusion. In order to eliminate noise, the facial area is segmented into the positive and negative areas as shown in Fig.2(a). The positive areas consist of real facial patches that contribute to detect facial features, while the negative areas may introduce errors to the task. The process of positive facial area extraction is given in Fig.2(b). In order to model the RF, patches from the positive facial area are labelled as $k = 1$ and others are labelled as $k = 0$. The training and testing are similar to RF [8, 12]. When a test patch P arrives at leaves of trees in the forest, we use the probability $p(c = k|l_t(P))$ stored at a leaf to judge whether the test patch belongs to the positive facial area. Only the patches from positive facial area are used for feature detection.

## 3. D-RF FOR FACIAL FEATURE DETECTION

In this section, the Dirichlet-tree distribution is introduced into RF framework to detect facial features. As shown in Fig.1, D-RF in-

*This work was supported by the National Key Technology Research and Development Program (No.2013BAH72B01) and research funds from Ministry of Education and China Mobile (MCM20130601), Research Funds of CCNU from the Colleges Basic Research and Operation of MOE (CCNU13B001), Wuhan Chenguang Project (2013070104010019), Central China Normal University Research Start-up funding (No.: 120005030223) , the Scientific Research Foundation for the Returned Overseas Chinese Scholars (No.:(2013)693), Young foundation of WenHua colloge(J0200540102), National Natural Science Foundation of China(No.61272206).




(a) (b)

**Fig. 2**. The positive facial area extraction.

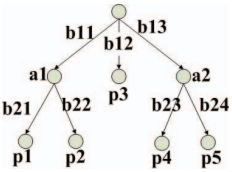



**Fig. 3**. A general Dirichlet-tree.

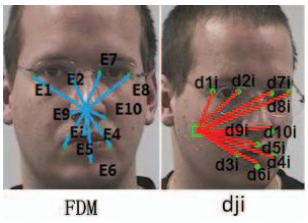



**Fig. 4**. Geometry features.

cludes two cascaded layers. First, 25 cascaded head pose models are estimated by D-RF in D-f1 as shown in Fig.5(a). Then, facial features are detected in local sub-regions which are found using Adaboost with Haar-like features in D-f2 as shown in Fig.5(b). Details on the D-RF are discussed in the following sections.

### 3.1. D-RF

The Dirichlet-tree is the distribution over leaf probabilities $[p_1 ... p_i]$ that results from this prior node probabilities $[a_1 ... a_k]$ on branch probabilities $b_{ji}$ [9], where $i$ is the number of a leaf , $k$ is the number of a prior node, $j$ is the layer of a branch as shown in Figure 3. RF is an ensemble approach in which several tree predictors are combined together to obtain high performance for classification or regression [12]. Each tree in RF is independently generated with random samples selected from the whole data set.

The D-RF arranges random trees of RF as the Dirichlet-tree structure as shown in Fig.3, where each node $a_f$ is a sub-forest. It is noted that each sub-forest is related to his prior node. Hence, the D-RF only computes final probabilities $p_i$ under its prior sub-forests instead of all trees probabilities in RF. Therefore, D-RF can provide high accuracy and efficiency.

**Training:** Each tree $T$ of a sub-forest in the D-RF $T = \{T_t\}$ is built and selected randomly from a different set of the training images. From each patch of positive areas, we extract combination feature sets $P_i = \{X_i, D_i|S_i, a_f\}$. $X_i = (x_{i1}, x_{i2}, x_{i3})$. represent multiple texture features which include Gabor features, LBP and gray values of the patch. $D_i = (d_{ji}, E_j|S_i, a_f)$ represent geometry features based on facial feature points and FDM, where $d_{ji}$ denote the $N$ 2D displacement vectors from the centroid of the patch $P_i$ to each facial feature points $n_j$, and $E_j$ denote the $N$ vectors from each facial point $n_j$ to the facial center point $F$ (see Fig.4 and Fig.5(c)), where $N$ is the number of facial points.

$$d_{ji} = ||n_j - P_i||_2, E_j = ||n_j - F||_2, j = 1, 2...N \quad (1)$$

$S_i$ is the head pose model that has been estimated previously. $a_f$ is the sub-forest in the node of D-RF.

In order to train the D-RF, we define a patch comparison feature as binary tests $\varphi$, similar to [2, 3]:

$$\varphi = |R_1|^{-1} \sum_{j \in R_1} X_y(j) - |R_2|^{-1} \sum_{j \in R_2} X_y(j) \quad (2)$$

where $R_1$ and $R_2$ are two random rectangles within the positive facial patches, $X_y(j)$ is the texture feature channel.

Then selecting a splitting candidate that best splits the feature set $P$ into two subsets $P_L$ and $P_R$, it maximizes the evaluation function Information Gain(IG),

$$P_L = \{P|\varphi < \tau\}, P_R = \{P|\varphi > \tau\} \quad (3)$$

$$IG = \arg\max_{\varphi}(H(P|S_i, a_f) - \sum\nolimits_{s \in \{L,R\}} \frac{|P_s|}{|P|} H(P_s|S_i, a_f)) \quad (4)$$

Where $\tau$ is threshold, $H(P|S_i, a_f)$ is the defined class uncertainty measure.

$$H(P|S_i, a_f) = -\sum_{i=1}^{N} \frac{\sum_i p(D_i|S_i, a_f, P_n)}{|P|} \log \left(\frac{\sum_i p(D_i|S_i, a_f, P_n)}{|P|}\right) \quad (5)$$

Where $p(D_i|S_i, a_f, P_n)$ indicates the probability that the positive patch $P_n$ belongs to the facial feature point $i$ in the sub-forest $a_f$ of the D-RF under the $S_i$ head pose.

Create leaf $l$ when IG is below a predefined threshold or when a maximum depth is reached. A leaf of D-RF for facial feature detection includes four learned models (see Fig.5(c)): (1)a positive/negative patch probability, (2)a head pose probability, (3) a probabilistic regression model for the locations of the base facial points, (4) a FDM model.

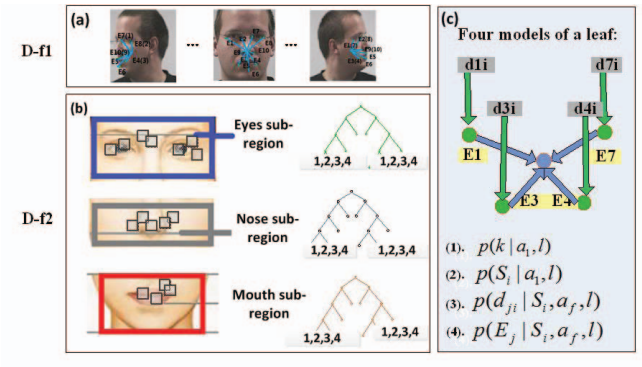



**Fig. 5**. The layers and leaf models of the D-RF. (a) Head pose estimation in D-f1. (b) The local sub-regions found using AdaBoost for local facial feature detection in D-f2. (c) Models of a Leaf.

During testing, $S_i$ and $a_f$ are estimated, then voting model with the estimated state is used. Details on the testing in each layer are given in the following sections

### 3.2. Head pose estimation in D-f1

Head pose estimation using the D-RF is similar to [6], where Liu *et al*. have proved its high accuracy and efficiency.

In leaves of four cascaded sub-layers from D-f1, there are 25 probabilitic models of head poses. The final head pose parameter is obtained by an adaptive Gaussian mixture model (GMM)[6, 13],

$$p(S_i^m|l_{b_i}) = N(S_i^m|b_i; \overline{S_i^m|b_i}, \sum{}_{b_i}) \quad (6)$$

Where $\overline{S_i^m|b_i}$ and $\sum{}_{bi}$ are the mean and covariance matrix of the head pose probabilities of the sub-forests $b_i$ in four cascaded sub-layers from D-f1.

In the case, we estimate 25 discrete yaw and pitch angles that are stored at leaves of D-f1, i.e. $\{90^\circ, 90^\circ\}, \{90^\circ, 45^\circ\}...\{-90^\circ, -90^\circ\}$. Our final head pose estimator reaches an accuracy of 71.83% on 25 head pose classification.

### 3.3. Facial features detection in D-f2

In D-f2, we detect facial features in different sub-regions (see Fig.5(b)). The algorithm pseudocode is shown in Fig.6.

**Algorithm 1 Facial feature detection**

**Input:** facial patches from a positive face area
**Output:** the 10 facial feature points.
***Step1*** *Loading relative sub-forests of the D-RF based on the head pose parameter in D-f1;*
***Step2****Locating the mouth, nose and eyes sub-region by AdaBoost;*
***Step3*** *Selecting patches around each local sub-region;*
***Step4*** *Loading sub-forests in D-f2 and testing for facial feature;*
***Step5*** *Voting leaf models to obtain 10 facial features.*

**Fig. 6**. The pseudocode of the D-RF for facial feature detection

Patches from local sub-regions are allowed to predict the locations of local points. It reduces the influences due to different poses, face deformation and local features. We measure the confidence $pf$ of a patch $P$ for the location of a feature point $j$ using,

$$pf \propto exp(\frac{||\overline{d_{ji}|k, S_i, a_f}||^2}{\gamma}) \bullet exp(\frac{||\overline{E_j|k, S_i, a_f}||^2}{\gamma}) \quad (7)$$

The constant $\gamma$ is used to control the steepness of this function. A patch with a high confidence $pf$ is only allowed to vote for feature points. The probabilities stored in the sub-forest $a_f$ from D-f2 is modeled as,

$$p(d_{ji}, E_j|S_i, A_f, P) = \frac{1}{T_f} \sum_i \sum_{t=1}^{k_f} p(d_{ji}, E_j|l_{t,a_f,S_i}(P)) \quad (8)$$

where $l_{t,a_f,S_i}(P)$ is the leaf model in the tree $T_f$ under the head pose $S_i$. The $K_f$ is the number of trees of sub-forests of D-f2. A leaf model should be learned if $E_j$ is under the predefined FDM,

$$p(d_{ji}, E_j|l_{t,a_f,s_i}(P)) = N(d_{ji}, E_j|a_f, S_i; \overline{d_{ji}, E_j|a_f, S_i}, \sum{}_{(d_{ji},E_j)l}) \quad (9)$$

where $\overline{d_{ji}, E_j|a_f, S_i}$ and $\sum{}_{d_{ji},E_j}$ are the mean value and covariance matrix of the offsets of the $j$-th facial feature point.

### 3.4. The composite weighted voting method

We use a composite vote method in a cascaded way. Both classification and regression method are used. In order to eliminate imbalance of samples in different subsets, we store the weight $w_s = P_s/P$ that is defined as the ratio of samples in a subset $P_s$ to full samples number $P$ in each tree of the D-RF. If voting for the $j$-th feature point is $D_{y_i}(j|a_f, S, k)$ in the patch location $y_i$, then we set the weighted voting model to be given as $V(j) \propto K((w_s D_{y_i} - (y_i + \overline{w_s D_{y_i}})/h_j)$. A Gaussian Kernel $K$ and the bandwidth parameter $h_i$ are given by GMM. Where $k = 1$ represents the positive facial patch, S={yaw, pitch} represents the classified head pose, then regression voting in local sub-regions $a_f$ can obtain good results by testing on sparse patches from cascaded sub-forests rather than all forests. Meanwhile, the competing method is casting GMM that is similar to [6]. The final 10 facial feature points are obtained by performing mean-shift in $V(j)$ for each point $j$.

## 4. EXPERIMENTS

Experimental results in unconstrained environment are given in this section, including Pointing04 head pose database [14] and LFW database [15]. The Pointing04 database consists of 2940 images with different poses and expressions. The LFW database consists of 5749 individual facial images and 13300 images, which have been collected in the wild and vary in poses, lighting conditions, resolutions, races, occlusions, and make-up. For evaluation, we divided the databases into a training set and a testing set. The training set consists of 2100 images from Pointing04 database and 12300 images from LWF database. The testing set includes the rest of 840 images from Pointing04 database, and 1000 images from LFW database.

### 4.1. Evaluation methodology

We measure the localization performance using the inter-ocular distance (IOD) normalized error, definite $e_i$ as localization error,

$$e_i = \frac{||I_i^G - I_i^D||_2}{I_{IOD}} \quad (10)$$

where $I_i^G$ is the ground truth location of the $i$-th point, $I_i^D$ is the detected location of the $i$-th point, and $I_{IOD}$ is the inter-ocular distance, which is defined as $|I_{LeftEye}^G - I_{RightEye}^G|$. A point is correctly detected if $e_i$ is below 0.1 IOD.

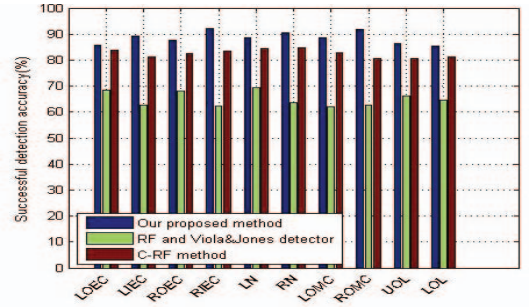
### 4.2. Experimental Results analysis

#### *4.2.1. Accuracy of the detected facial points*

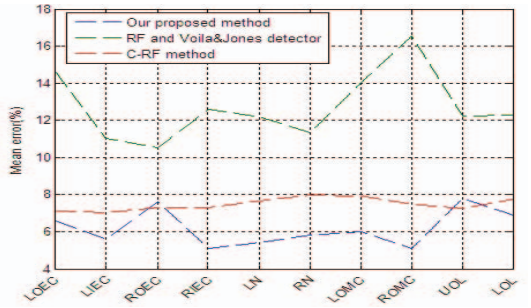
The key setting parameters on testing and training of the D-RF include: the maximum depth of each tree (i.e.,15), test candidates at split node (i.e.,2500), face box size ((i.e.,125*125 pixels), patch size((i.e.,0.25*face box size), the maximize offset distant between facial points and center location of random facial patches (40), the number of mean-shift iterations (7), the bandwidth of the mean-shift kernel (10). Table 1 shows the detection rate of each facial point using the proposed approach.

#### *4.2.2. Comparison with state of the Art*

In Fig.7, we compare the D-RF with state-of-the-art approaches on the same databases, i.e., C-RF [3] and RF+Viola&Jones method [11, 12]. One can see that our approach performs better than C-RF and RF +Viola &Jones. Additionally, some examples of detection results are shown in Fig.8. In some wide range head pose variations, occlusion and illuminations, our approach performs well.




(a)Detection accuracies




(b) Mean errors

**Fig. 7**. Comparison with state of the Art.

**Table 1**. Detection rate of each facial point in the databases (%)

| Facial feature point | Accuray | Mean error |
|---|---|---|
| **Left outside eye corner(LOEC)** | 85.6 | 6.4 |
| **Left inner eye corner(LIEC)** | 89.3 | 5.6 |
| **Right outside eye corner(ROEC)** | 87.7 | 7.3 |
| **Right inner eye corner(RIEC)** | 92.2 | 5.1 |
| **Left nostril(LN)** | 88.5 | 5.4 |
| **Right nostril(RN)** | 90.7 | 5.7 |
| **Left outside mouth corner(LOMC)** | 88.6 | 6.2 |
| **Right outside mouth corner(ROMC)** | 91.7 | 5.3 |
| **Upper outside lip(UOL)** | 86.4 | 7.5 |
| **Lower outside lip(LOL)** | 85.5 | 6.8 |

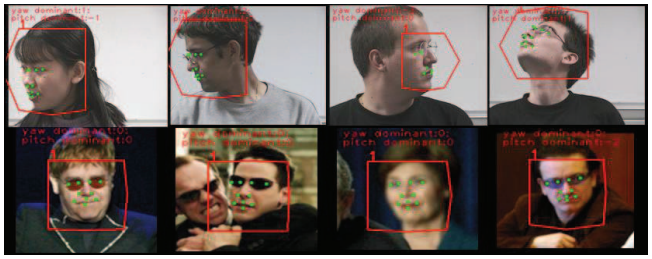


**Fig. 8**. Examples of detection results using the D-RF.

*4.2.3. Comparison between the D-RF with FDM and the D-RF without FDM*

To evaluate the efficiency of FDM, the average detection accuracies of the D-RF with FDM and without FDM are given in Fig.9. The comparison results show that the D-RF with FDM provides higher accuracies than the D-RF without FDM under the 25 head pose models, particularly, under the head poses of large rotation angles.

*4.2.4. Computation time*

The experiments have been conducted on a PC with Intel(R) Core(TM) i5-2400 CPU@ 3.10GHz. The computation time of D-RF, C-RF and RF is given in Table 2. From the table, one can see that the D-RF is faster than the C-RF and RF.

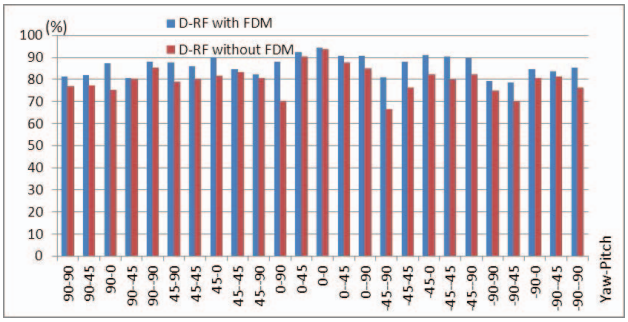



**Fig. 9**. D-RF with FDM vs. D-RF without FDM.

**Table 2**. Computation time (/s) in the D-RF, C-RF and RF.

| Algorithm | D-f1 | D-f2 | Total |
|---|---|---|---|
| D-RF | 0.4015 | 0.25609 | 0.65709 |
| C-RF | — | | 0.76335 |
| RF | — | | 1.04723 |

## 5. CONCLUTIONS

In this paper, we propose a cascaded learning approach for facial feature detection in unconstrained environment. First the positive facial area is extracted to eliminate the influence of noise in unconstrained environment. Then D-RF is proposed to detect facial features under 25 cascaded head pose models, local sub-regions and FDM. Furthermore, multiple probability models are learned and stored in leaves of the D-RF. Finally, the composite weighted voting method is used to vote multiple models stored in leaves and eliminates influence of samples in imbalance distribution. Experiment results with two public databases demonstrate the robustness and accuracy of the proposed approach. In future work, this method could be extended to detect vision attention in a wide scene, e.g. the students attention in a classroom.

## 6. REFERENCES

[1] T.F. Cootes, M.C. Ionita, and S.P., "Robust and accurate shape model fitting using random forest regression voting," in *Proc. European Conf.* Computer Vision, 2012.

[2] H. Yang and I. Patras, "Face parts localization using structured-

output regression forests," in *Proc. Assian Conf*. Computer Vision. Springer, 2012.

[3] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real time facial feature detection using conditional regression forests," in *CVPR*, 2012.

[4] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *TPAMI*, 2001, vol. 23, pp. 681–685.

[5] H. Yang and I. Patras, "Privileged information-based conditional regression forests for facial feature detection," in *Proc. IEEE Int1 Conf*. on Automatic Face and Gesture Recognition, 2013.

[6] Y. Liu, J. Chen, Y. Liu, Y. Gong, and N. Luo, "Dirichlet-tree distribution enhanced random forests for head pose estimation," in *ICPRAM*, 2014.

[7] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," in *Maching Learning*, 1999, vol. 37(3), pp. 297–336.

[8] Y. Li, X. Wang, and X. Ding, "Person-independent head pose estimation based on random forest regression," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1521–1524.

[9] T. Minka, "The dirichlet-tree distribution," in *http://research.microsoft.com/minka/papers/dirichlet/minkadirtree.pdf*, 1999.

[10] X. Yan and C. Han, "Mutiple target tracking by probability hypothesis density based on dirichlet distribution," in *Journal of XiAn JiaoTong University*, 2011, vol. 42, p. 2.

[11] M. E. Jones and P. Viola, "Fast multi-view face detection," in *Tech. Rep. TR2003-096*. Mitsubishi Electric Research Laboratories, 2003.

[12] L. Breiman, "Random forests," in *Machine Learning*, 2001, vol. 45(1), pp. 5–32.

[13] C. E. Stauffer and W. Grimson et al, "Adaptive background mixture models for real-time tracking[c]," in *Computer Vision and Pattern Recognition*, 1999.

[14] N. Gourier, D. Hall, and J. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Pointing*. ICPR international Workshop on Visual Observation of Deictic Gestures, 2004, pp. 183–191.

[15] G. Huang, M. Ramesh, T. Berg, and E., "Learned-miller. labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Technical report, University of Massachusetts, Technical report, University of Massachusetts*, 2007.