Solar Physics (2006) 237: 419–431
DOI: 10.1007/s11207-006-0114-5

AUTOMATIC DETECTION AND CLASSIFICATION OF CORONAL MASS EJECTIONS

MING QU and FRANK Y. SHIH
College of Computing Science, New Jersey Institute of Technology, Newark, NJ 07102, U.S.A.
(e-mail: shih@njit.eu)

and

JU JING and HAIMIN WANG
Center for Solar-Terrestrial Research, New Jersey Institute of Technology Newark, NJ 07102, U.S.A.;
Big Bear Solar Observatory, New Jersey Institute of Technology, 40386 North Shore Lane, Big Bear City, CA 92314, U.S.A.

(Received 1 December 2005; accepted 2 July 2006; Published online 27 August 2006)

Abstract. We present an automatic algorithm to detect, characterize, and classify coronal mass ejections (CMEs) in Large Angle Spectrometric Coronagraph (LASCO) C2 and C3 images. The algorithm includes three steps: (1) production running difference images of LASCO C2 and C3; (2) characterization of properties of CMEs such as intensity, height, angular width of span, and speed, and (3) classification of strong, median, and weak CMEs on the basis of CME characterization. In this work, image enhancement, segmentation, and morphological methods are used to detect and characterize CME regions. In addition, Support Vector Machine (SVM) classifiers are incorporated with the CME properties to distinguish strong CMEs from other weak CMEs. The real-time CME detection and classification results are recorded in a database to be available to the public. Comparing the two available CME catalogs, SOHO/LASCO and CACTus CME catalogs, we have achieved accurate and fast detection of strong CMEs and most of weak CMEs.

1. Introduction

Coronal mass ejections (CMEs) are perhaps the most important solar energetic events as far as space weather is concerned. The Large Angle Spectrometric Coronagraph (LASCO) on aboard the Solar and Heliospheric Observatory (SOHO) spacecraft has been in operation since 1996 and has produced thousands of coronal images (Brueckner *et al.*, 1995). LASCO C1, C2, and C3 images, each tailored to a specific field-of-view of 3, 6, and 30 solar radii, respectively, provide useful tools to observe CMEs at their early stage.

The structures of CMEs can be grossly arranged into eight categories (Howard *et al.*, 1985): halo, curved front, loop, spike, double spike, multiple spike, streamer blowout, diffuse fan, and complex. In an effort to get an intensity classification, Dai, Zong, and Tang (2002) analyzed the velocity, span, mass, and kinetic energy of CMEs and proposed three CME intensity categories, strong (including halo, complex), middle (including double spike, multiple spike, and loop), and weak (including spike, streamer blowout, and diffuse fan). Among these categories, the halo CMEs, which appear as expanding, circular brightenings surrounding the coronagraph's occulter, are of great concern because they may carry away a mass of up to 10^{15} kg and are moving outward along the Sun–Earth line with speeds greater than 500 km s^{-1} (Dai, Zong, and Tang, 2002). It appears that halo CMEs are an excellent indicator of increased geo-activity 3–5 days later (Brueckner *et al.*, 1998), and therefore, the automatic and accurate detection and characterization of CMEs, especially halo CMEs, in real-time can provide an early warning of the occurrence of potentially geo-effective disturbance. In Figure 1, examples of a strong and weak CMEs in LASCO C2 images are pointed by arrows.

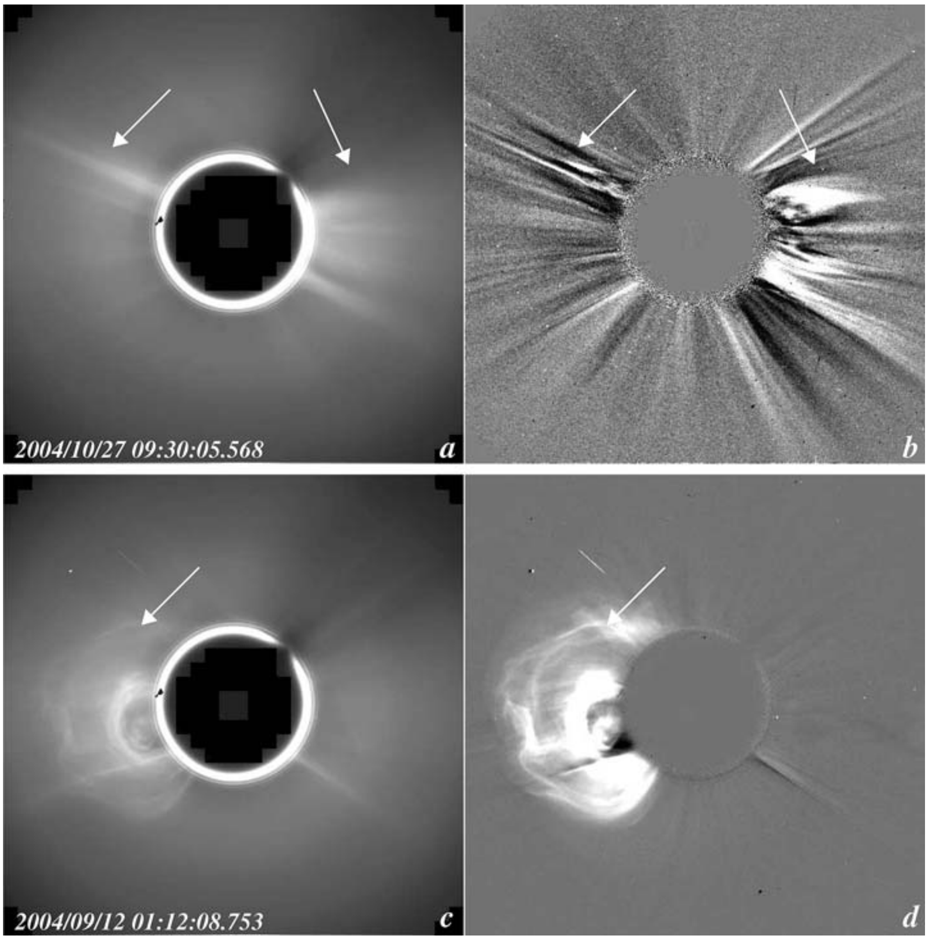

Figure 1. (a) Weak CMEs in a LASCO C2 image on October 27, 2004. (b) Same weak CMEs in the running difference image. (c) A strong CME in a LASCO C2 image on September 12, 2004. (d) Same strong CME in the running difference image. CMEs are pointed by *arrows*.

The traditional way of detecting CMEs is based on human observation. For example, the SOHO/LASCO group provided a CME catalog based on image enhancement and human detection. However, the human observation is subjective and slow, and decision rules may vary from one operator to another. During the last few years, image processing and pattern recognition techniques have been utilized in automatic CME detection. For example, Berghmans (2002) introduced an automatic CME detection procedure using image processing techniques such as image enhancement, thresholding, and motion tracking. In 2004, Robbrecht and Berghmans (2004) developed an application, computer-aided CME tracking (CACTus), for automatic detection of CMEs in sequences of LASCO images which uses Hough transform to detect CMEs in the running differences. They also published their real-time CME detection results through the Internet.

In this paper, we present an automatic three-step CME detection and classification algorithm. In brief, the three steps are as follows: (1) preprocessing, in which running difference images of LASCO C2 and C3 are automatic generated; (2) detection and characterization, in which the structural properties of CMEs (such as height, span angle) are characterized using image segmentation and morphological methods; and (3) classification, in which SVM classifiers are used to classify CMEs into three groups of strong, medium, and weak on the basis of characterization. Three steps are described in Sections 2–4, respectively. Finally, a comparison of LASCO, CACTus, and our CME catalogs is presented.

2. Preprocessing

SOHO/LASCO provides C2 and C3 images in a Flexible Image Transport System (FITS) format which can be downloaded from the SOHO/LASCO web site. Image information such as size, center coordinates, and exposure time is written into the headers of FITS images. In our preprocessing step, LASCO images in different sizes are resized to the same size of 512 pixels × 512 pixels and aligned according to coordinates of solar centers. Like most other CME detections, brightness normalization is applied according to the ratio of exposure time of LASCO images. Furthermore, if the gray level ranges of mean brightness between two consecutive images are different after the exposure time adjustment, the images are normalized using the mean brightness ratio of the current image to the previous image. In our preprocessing, when the ratio of their gray level is greater than 1.1 or less than 0.9, Equation (1) is applied to correct the brightness of the current image.

$$M'_{\mathrm{c}} = M_{\mathrm{c}} \frac{\bar{x}_{\mathrm{p}}}{\bar{x}_{\mathrm{c}}}, \tag{1}$$

where M_{c} and M'_{c} are the current images before and after the normalization, and $\bar{x}_{\mathrm{p}}$ and $\bar{x}_{\mathrm{c}}$ are the mean brightness of no-CME regions on the previous and current images, respectively. The mean brightness ratio is based on the regions with no

CME because the mean brightness normalization may reduce CMEs' brightness. The pixel with a gray level lower than the median value of the image is considered to be the one located on the no-CME region.

In LASCO images, streamers are structured similarly to CMEs. The difference between streamers and CMEs are that streamers are steady bright structures, while CMEs are sudden brightness changes. To distinguish a CME from other stable bright structures (*e.g.*, streamers) and stable bright noise, a running difference image, obtained by Equation (2), is commonly used. The running difference image is also useful for the measurement of CMEs' moving-front (*i.e.*, the top moving edge) and foot-point (*i.e.*, the lower moving bottom).

$$D = M_{\mathrm{c}} - G, \tag{2}$$

where G is the reference image and M_{c} the current LASCO image. The reference image to use with the normal aspect is the image before the current image. This has the disadvantage that if some regions of the previous image are badly captured, the resulting running difference images are badly generated. Another issue is when a CME is evolved on a sequence of image frames, the running difference may succeed in detecting the moving front of the CME, but may fail in detecting its location of foot-point because some parts of CME regions are overlapped on a sequence of images. In this paper, a reference image G' is produced recursively by using the combination of the reference and previous images, as shown in Equation (3).

$$G' = Gc_1 + M_{\mathrm{c}}c_2, \tag{3}$$

where c_1 and c_2 denote the percentages of effectiveness on the reference and current images. The disadvantage of using a single image as the reference image is avoided. Based on our experiments, we set $c_1 = 90\%$ and $c_2 = 10\%$. However, if a strong CME is detected in the previous image frame, $c_1 = 100\%$ and $c_2 = 0\%$ are used.

In the next section, CMEs are segmented using a threshold based on high brightness regions. When a weak CME with small brightness enhancement is located in the dark region and a strong CME is located nearby, the weak CME may not be detected because the chosen threshold is focused on the strong one. To overcome this problem, a partial division image is used to segment CME regions based on the ratio of brightness between the current LASCO image and reference image. The division image is obtained by dividing the LASCO image by the reference image. For the region with gray level close to zero, the division may enhance noise significantly and produce overflow errors. Therefore, only a partial region of the image is chosen. The partial division image V is shown in Equation (4).

$$V = \frac{M_2}{G_2}, \tag{4}$$

where G_2 is the region on the reference image with pixels brighter than the median value of the reference image and M_2 the region corresponding to G_2 on the LASCO image. By using the relatively bright regions for G_2, the division over flow is avoided.

Some of LASCO images contain missing blocks, which appear as dark or bright squares on image frames. Three criteria are proposed to find the missing blocks: (1) when a dark missing block first appears in the current image, its appearance on the current image and the running difference images are both dark; (2) its gray level of the current image is close to the minimum value of the image, and its gray level of the running difference image is less than a negative threshold t_1; (3) the number of pixels for an individual block region is set to be greater than a threshold t_2 because a missing block contains a large amount of pixels as compared to small noise. The thresholds t_1 and t_2, obtained from our experiments, are negative values of the standard deviation of the running difference in brightness and 50 pixels, respectively. A bright missing block can also be detected in the same way. Finally, those missing blocks in running difference images and division images are replaced by the minimum value of the images.

3. Automatic Detection of CMEs

Two consecutive LASCO images incorporated with the reference image produce two running difference images and two division images. The running difference images and division images are segmented into binary differences by applying a thresholding method. By tracking segmented CME regions in binary differences, the properties of CMEs are obtained.

3.1. Segmentation of CMEs

Running difference images (D) and division images (V) are constructed by using consecutive LASCO images in the previous step. In this step, a thresholding method is adopted to D and V to produce binary differences in which CME regions are separated from the background. A fixed value of the threshold would produce unstable results because contrasts of running difference images vary from time to time. In our CME segmentation, the threshold is computed using the median and standard deviation of D and V. Based on extensive experiments, the thresholds for D and V of C2 images are chosen as $m + s$, where m and s are the median and standard deviation of C2's D and V images, respectively. Similarly, we obtain the thresholds for C3's D and V images which is $m + 1.5s$. These automatic thresholds are robust with respect to different image contrasts. The final segmentation result is the summation of the two segmentation results from the difference and the division images.

After segmentation, morphological closing (Shih and Mitchell, 1989) is applied to the binary segmentation image to eliminate small gaps. A 5 × 5 structuring element is used to perform the binary closing. The closing of a region A by a structure element B, denoted by $A \bullet B$, is defined as

$$A \bullet B = (A \oplus B) \ominus B. \tag{5}$$

The closing of A by B is simply the dilation of A by B, followed by the erosion of the result by B. The closing method can smooth features and remove small noise. The 5 × 5 structuring element removes the noise whose width and length are less than 5 pixels and eliminates the gaps whose distance is less than 5 pixels.

In order to calculate the features of CMEs easily, binary segmentation images are reformed to $360r$ binary angular images, where r is the radius to the edge of the C2 or C3 occulting disk in the LASCO image. The results of the segmentation, morphological closing, and the angular images are shown in Figure 2. The degree of angle $[0, 359]$ is counted clockwise from north. The sizes of C2 and C3 angular images are 360 pixels × 156 pixels and 360 pixels × 216 pixels, respectively.

3.2. Features of CMEs

In the feature detection, all the segmented regions in a CME frame become CME candidate regions, but only continuously moving regions on consecutive images are

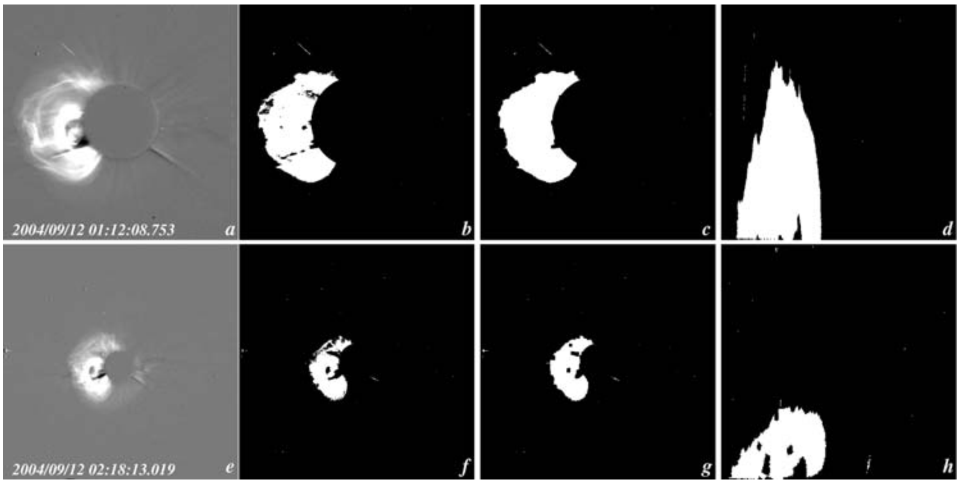

Figure 2. (a) The running difference of a LASCO C2 image observed on September 12, 2004. (b) Binary segmented result of the LASCO C2. (c) Morphological closing result of the LASCO C2. (d) Angular image of the LASCO C2. (e)–(h) The running difference of a LASCO C3 image and its segmented, closing, and angular results, respectively.

treated as real CME regions. A CME candidate region is classified as a CME region if it occurs on the current image frame as well as on the previous image frame. The corresponding rules are given as follows: (1) the distance between the two centers of overlapping regions is less than t_3(time interval/1 h), where t_3 is the threshold for the maximum CME movement for an hour; and (2) the difference between the span widths of two overlapping regions is less than t_4(time interval/1 h), where t_4 is the threshold for the maximum degree span change for an hour. t_3 is chosen as 100 pixels for the LASCO C2 image and as 40 pixels for C3 image. t_4 is chosen as a half of the span width of the CME candidate. In addition, the speed and the new increasing area for CME regions are computed by comparing the corresponding CME regions on the two binary angular images. Let us denote a CME region in the current image and its corresponding region in the previous image as A and A_p, respectively. The CME properties of A in an image frame are listed in Table I.

TABLE I

The properties of a CME region.

No.	Description of the CME properties
1	The exposure time of the LASCO image
2	The time interval between the current and the previous image
3	The pixel size of the LASCO image
4	The mean brightness value of the reference image
5	The mean brightness value of the current image
6	The mean brightness value of the running difference
7	The standard deviation of the running difference
8	The number of pixels for A
9	The threshold for segmenting A from the running difference
10	The maximum height (arcsecs from disk center) of A
11	The height of the center of A
12	The minimum height of A
13	The starting angle of A The angle is calculated from North 0 clockwise
14	The angle of the center of A
15	The ending angle of A
16	The angular width of A
17	The height difference (h_1) between the maximum height of A and A_p
18	The height of the new moving region (h_2) which is obtained by subtracting A_p from A
19	The speed which is computed using h_1, divided by the interval time cadence
20	The speed which is computed using h_2 divided by the interval time cadence
21	The span width of the new moving region
22	The center angle of the new moving region

Features 17–22 are obtained using the corresponding regions on the two consecutive images.

The detection reports for a CME on August 7, 2005 are shown in Figures 3 and 4. The classification of strong, medium, and weak CMEs is presented in the next section. Figures 5 and 6 show the height and velocity for the CMEs on September 1, 2002 in C2 and C3 images, respectively. The speed of the CME is around 300 km s^{-1}, which is almost equal to the computed CME speed in the LASCO catalog.

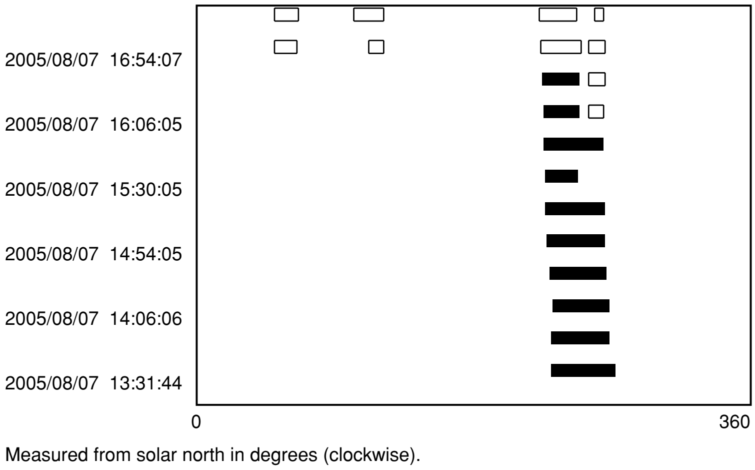

Figure 3. Detection and classification report for a detected CME using LASCO C2 images on August 7, 2005. Apparently, there is a strong CME on the west from 13:30 to 16:50 UT. *Solid black* and *empty rectangles* denote strong and weak CMEs, respectively.

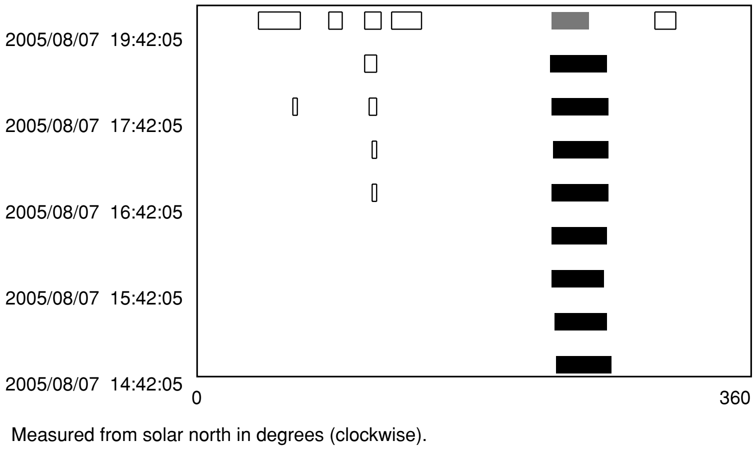

Figure 4. Detection and classification report for a detected CME using LASCO C3 images on August 7, 2005. There is a strong CME on the west from 14:40 to 19:40 UT. *Solid black*, *solid gray*, and *empty rectangles* denote strong, medium, and weak CMEs, respectively.

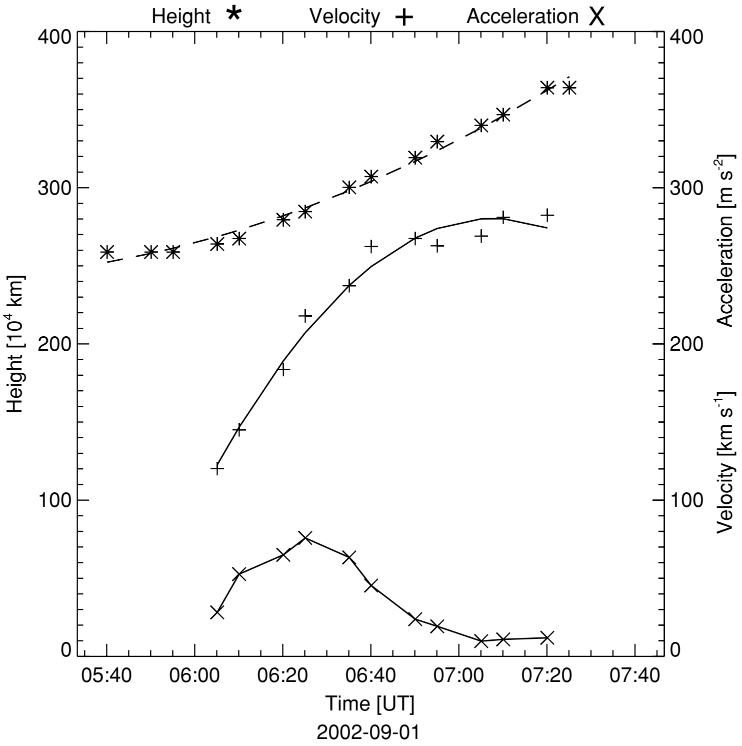

Figure 5. Height, velocity, and acceleration profile of a CME for LASCO C2 images on September 1, 2002.

4. Classification of Strong, Medium, and Weak CMEs

A strong CME consists of a large amount of fast moving mass. In our CME classification step, strong CMEs are selected from halo, curved front, and complex CMEs according to Howard *et al.* (1985). In this paper, the Support Vector Machine (SVM) classifier with a linear kernel is used for distinguishing the strong CME from others. For the present study, a CME can be represented by 22 features, which are obtained in the previous step. The SVM training computes the corresponding weights for each input feature for the CME classification. Six features with significant weights are selected as the inputs for classification. The six input features are proven to be robust based on our comparisons on feature combination.

The SVM is a powerful learning system for data classification. The idea of SVMs is to divide the fixed given input pattern vectors into two classes using a hyperplane with the maximal margin, which is the distance between the decision plane and the closest sample points (Vapnik, 1998). When support vectors (the points on the decision plane) are found, the problem is solved. Support vectors on the decision plane can be found by the classical method of Lagrange multipliers.

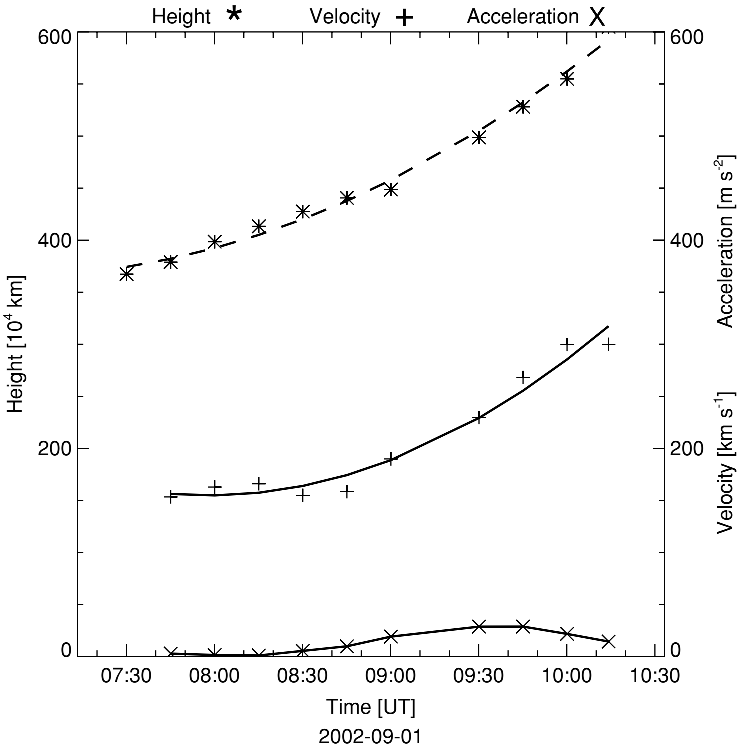

Figure 6. Height, velocity, and acceleration profile of a CME for LASCO C3 images on September 1, 2002.

Compared to other learning machine such as neural networks based on the empirical risk minimization (ERM), SVM is based on the structural risk minimization (SRM) inductive principle. The SRM principle is intended to minimize the risk functional with respect to the empirical risks in the ERM and the confidence interval (Vapnik, 1998). Because SVM considers both terms in the risk function, its classifier is better than neural networks, which only consider the empirical risks. For the nonlinear case, SVM uses the kernel functions to transfer input data to feature space. For example, polynomial support vector classifier and Gaussian radial basis function (RBF) kernel classifier are two popular nonlinear classifiers. In our study, a simple linear SVM is used. To increase the complexity of classification, the nonlinear kernel could be used in the future study. This is the first time that the SVM is applied to the CME classification. The comparisons between the linear and other kernels can be found in Qu *et al.* (2003). For further information regarding the SVM, readers can refer to Vapnik (1998).

The six inputs to our SVM classifier are the features with significant classification weights in Table I which are as follows: the mean brightness in the running difference image, the number of pixels in the running difference image,

TABLE II

Success classification rates of the strong and non-strong CMEs based on 50 strong CMEs and 50 non-strong CMEs.

LASCO	Strong CMEs (%)	Weak and medium CMEs (%)
C2	94	96
C3	96	96

We assume the detection by human operators to be 100% accurate.

the angular width of span, the height of new moving region, the span width of the new moving region, and the speed described in feature 20 of Table I. A list of CMEs are randomly selected in 2004. Assuming that human classification for strong CMEs is 100% accurate, we can select 50 strong (halo, curved front, and most complex) and 50 other (medium and weak) CMEs by searching through a sequence of running difference images. The SVM classifier is trained by human classification results with the 100 CMEs based on the aforementioned six inputs. After training, the SVM classifier is able to classify strong CMEs from others automatically. The classification rate of the testing experiments is shown in Table II.

After finding strong CMEs using our SVM classifier, further classification is applied to distinguish medium from weak CMEs using the rule proposed by Howard *et al.* (1985) and Dai, Zong, and Tang (2002). The speed of medium CMEs is greater than 300 km s^{-1}, while the speed of weak CMEs is less than 300 km s^{-1}.

5. Comparisons for CME Detections

We have developed the software to detect and characterize CMEs. The programs were developed in Interactive Data Language (IDL) by Research Systems Inc., and run on a DELL Dimension 4600 PC with CPU time 2.8 GHz and memory of 512 MB under Linux. The computational time for detecting a CME using three LASCO images is about 5 s, which is far less than the observational interval. The results in our catalog include a list of CMEs, a sequence of CME image frames, the classification type of each CME frame, and the properties such as height, velocity, and angular width of each CME region. The properties and classification results of CMEs are saved in our database available through our web site.

There are two, previously developed, CME detection catalogs available to the public. From the LASCO web site at *http://cdaw.gsfc.nasa.gov/ cme_list/*, one can find the CME catalog created by visual inspection. Robbrecht and Berghmans presented their results at *http://sidc.oma.be/ cactus/*. Our automatic detection results are currently shown at *http:// filament.njit.edu/detection/vso.html*. It is difficult to compare the CME catalogs because there is no comprehensive catalog to be used as a reference to conduct the comparisons (Berghmans, 2002). A CME in one catalog may be considered as two CMEs in another, and the beginning and ending time for a CME is hard to determine. The reasons are that the preprocessing methods and detection rules of three catalogs are different and human decision is subjective. The reference image and threshold selection may affect the final decision of a CME detection.

We select results between August 1 and 31, 2004 to perform comparisons among the three catalogs. The LASCO catalog, used as the reference, listed 65 CMEs in this period. Our catalog missed one weak CME which happened on the west at 316° on August 26, 2004 at 16:54 UT. CACTus missed three weak CMEs which happened on the east at 77° on August 2, 2004 at 23:06 UT, on the east at 94° on August 22, 2004 at 17:30 UT, and on the east at 111° on August 26, 2004 at 21:54 UT. The missed CMEs in our catalog and in CACTus are weak CMEs. On the other hand, some CMEs missed in LASCO can be detected in CACTus and our catalog. For example, the CMEs happened on the east on August 6, 2004 at 22:30 UT, on the east on August 7, 2004 at 18:54 UT, and on the southeast on August 9, 2004 at 21:30 UT. By combining the results of CACTus and our catalog, the CME detection results are more complete and accurate than the LASCO catalog which is based on human eye detection. In the three catalogs, the center of principal angles, angular width and speed of the strong CMEs are given as the important CME properties. The detected properties of CMEs in the different catalogs could be varied because the region merging criteria are different. As we said previously, a CME in one catalog may be counted as two in another. Overall, the center of principal angles, the angular width, and speed, especially for the strong CMEs, are similar in all the three catalogs.

6. Summary

In this paper, an automatic algorithm to detect and categorize CMEs is presented. The first preprocessing step intends to normalize the images, remove the missing blocks, and obtain the running differences. Two consecutive images accompanying the reference image are used to produce two running differences, two division, two binary, and two angular images. The properties of CMEs such as intensity, height, span, and velocity are measured using automatic thresholding and morphology methods. The detected CMEs are saved to our database and are listed through our web site. In addition, a new advanced method, SVM, is used to distinguish strong CMEs from other kinds of CMEs. The strong CME classification is essential to the forecast of space weather. Compared to the previous CME detection methods proposed by SOHO/LASCO and CACTus, we have provided an alternative approach for CME detection and have proposed a new method for strong CME classification.

Acknowledgements

This work is supported by National Science Foundation (NSF) under grants IIS-0324816, ATM 0548956, and ATM 0536921.

References

Berghmans, D.: 2002, *Proceedings of the 10th European Solar Physics Meeting SP-506*, **1**, 85.

Brueckner, G.E., Howard, R.A., Koomen, M.J., Korendyke, C.M., Michels, D.J., Moses, J.D., Socker, D.G., Dere, K.P., Lamy, P.L., Llebaria, A., Bout, M.V., Schwenn, R., Simnett, G.M., Bedford, D.K., and Eyles, C.J.: 1995, *Solar Phys.* **162**, 357.

Brueckner, G.E., Delaboudiniere, J.-P., Howard, R.A., Paswaters, S.E., St. Cyr, O.C., Schwenn, R., Lamy, P., Simnett, G.M., Thompson, B., and Wang, D.: 1998, *Geophys. Res. Lett.* **25**, 3019.

Dai, Y., Zong, W., and Tang, Y.: 2002, *Chin. Astron. Astrophys.* **26**, 183.

Howard, R.A., Sheeley, N.R., Jr., Michels, D.J., and Koomen, M.J.: 1985, *J. Geophys. Res.* **90**, 8173.

Qu, M., Shih, F.Y., Jing, J., and Wang, H.: 2003, *Solar Phys.* **217**, 157.

Robbrecht, E. and Berghmans, D.: 2004, *Astron. Astrophys.* **425**, 1097.

Shih, F.Y. and Mitchell, O.R.: 1989, *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 31.

Vapnik, N.V.: 1998, *Statistical Learning Theory*, Wiley, New York.